Multilevel weighted least squares polynomial approximation

ESAIM: M2AN 54 (2020) 649–677
https://doi.org/10.1051/m2an/2019045

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

MULTILEVEL WEIGHTED LEAST SQUARES POLYNOMIAL APPROXIMATION

Abdul-Lateef Haji-Ali[1], Fabio Nobile[2], Raúl Tempone[3,4] and Sören Wolfers[3,*]

Abstract. Weighted least squares polynomial approximation uses random samples to determine projections of functions onto spaces of polynomials. It has been shown that, using an optimal distribution of sample locations, the number of samples required to achieve quasi-optimal approximation in a given polynomial subspace scales, up to a logarithmic factor, linearly in the dimension of this space. However, in many applications, the computation of samples includes a numerical discretization error. Thus, obtaining polynomial approximations with a single level method can become prohibitively expensive, as it requires a sufficiently large number of samples, each computed with a sufficiently small discretization error. As a solution to this problem, we propose a multilevel method that utilizes samples computed with different accuracies and is able to match the accuracy of single-level approximations with reduced computational cost. We derive complexity bounds under certain assumptions about polynomial approximability and sample work. Furthermore, we propose an adaptive algorithm for situations where such assumptions cannot be verified *a priori*. Finally, we provide an efficient algorithm for the sampling from optimal distributions and an analysis of computationally favorable alternative distributions. Numerical experiments underscore the practical applicability of our method.

Mathematics Subject Classification. 41A10, 41A25, 41A63, 65B99, 65N22.

Received October 4, 2017. Accepted June 21, 2019.

1. Introduction

A common goal in uncertainty quantification [23] is the approximation of response surfaces

$$\boldsymbol{y} \mapsto f(\boldsymbol{y}) := Q(u_{\boldsymbol{y}}) \in \mathbb{R},$$

which describe how a quantity of interest Q of the solution $u_{\boldsymbol{y}}$ to some partial differential equation (PDE) depends on parameters $\boldsymbol{y} \in \Gamma \subset \mathbb{R}^d$ of the PDE. The non-intrusive approach to this problem is to evaluate the response surface for finitely many values of $\boldsymbol{y}$ and then to use an interpolation method, such as (tensor-)spline interpolation [9], kernel-based approximation (kriging) [15,31], or (global) polynomial approximation [23].

Keywords and phrases. Multilevel methods, least squares approximation, multivariate approximation, polynomial approximation, convergence rates, error analysis.

[1] Heriot-Watt University, Edinburgh, UK.

[2] École Polytechnique Fédérale de Lausanne (EPFL), CSQI-MATH, Lausanne, Switzerland.

[3] King Abdullah University of Science and Technology (KAUST), CEMSE, Thuwal, Saudi Arabia.

[4] RWTH Aachen University, Department of Mathematics, Aachen, Germany.

*Corresponding author: soeren.wolfers@kaust.edu.sa

In this work, we study a variant of polynomial approximation in which least squares projections onto finite-dimensional polynomial subspaces are computed using values of f at finitely many random locations. More specifically, given a probability measure μ on the parameter space Γ and a polynomial subspace $V \subset L^2_\mu(\Gamma)$, the approximating polynomial is determined as

$$\Pi_V f := \operatorname*{arg\,min}_{v \in V} \|f - v\|_N \,, \tag{1.1}$$

where $\|\cdot\|_N$ is a discrete approximation of the $L^2_\mu(\Gamma)$ norm that is based on evaluations in finitely many randomly chosen sample locations $\boldsymbol{y}_j \in \Gamma$, $j \in \{1, \dots, N\}$ and a weight function $w\colon \Gamma \to \mathbb{R}$

The case where equally weighted samples are drawn independently and identically distributed from the underlying probability measure itself, $\boldsymbol{y}_j \sim \mu$, has been popular among practitioners for a long time and has been given a thorough theoretical foundation in the past decade [4, 8, 28]. More recently, the use of alternative sampling distributions and non-constant weights was studied in [6, 18, 29]. In particular, Hampton and Doostan [18] presented a sampling distribution ν_V^* and a corresponding weight function for which the number of samples required to determine quasi-optimal approximations within V is bounded by $\dim V$ up to a logarithmic factor. (This result was proved in [18] for total degree polynomial spaces and generalized in [6] to more general function spaces.) Since this distribution depends on V, it is natural to ask how samples can be efficiently obtained from it and whether there is an alternative that works equally well for all polynomial subspaces V. To address the first question, we present and analyze an efficient algorithm to generate samples from ν_V^* in the case where Γ is a product domain and μ is a product measure. For more general cases, we also study Markov chain methods for sample generation and analyze the effect of small perturbations of the sampling distribution on the convergence estimates of [6, 18]. To address the second question, we provide upper and lower bounds on ν_V^* in the case where Γ is a hypercube. The lower bound allows us to make the error estimates obtained in [6] more explicit. The upper bound shows that the arcsine distribution, which was proposed in [29], performs just as well as ν_V^* up to a constant that is independent of V but increases exponentially as the dimension d of Γ increases.

To motivate the main contribution of this work, namely the multilevel weighted least squares polynomial approximation method, we note that the response surface f from the beginning of this introduction cannot be evaluated exactly. Indeed, in most cases, the computation of $Q(u_{\boldsymbol{y}})$ requires the numerical solution of a PDE. Thus, we can only compute approximations of f whose accuracy and computational work are determined by the PDE discretization. If we simply applied polynomial least squares approximation using a sufficiently fine discretization of the PDE for all evaluations, then we would quickly face prohibitively long runtimes. For this reason, we introduce a multilevel method that combines numerous cheap samples using coarse discretizations with relatively few more expensive samples using fine discretizations of the PDE. In the recent decade, such multilevel algorithms have been studied intensely for the approximation of expectations [16, 19, 21, 22]. The goal of this paper is to extend this earlier work to the reconstruction of the full response surface, using global polynomial approximation and estimating the resulting error in the L^2_μ norm.

To describe the multilevel method, assume that we want to approximate a function f. Assume furthermore that we can only evaluate functions f_l with $f_l \to f$ as $l \to \infty$ in a suitable sense and that the cost per evaluation increases as $l \to \infty$. A straightforward approach to this situation is to apply least squares approximation to some f_L that is sufficiently close to f. The theory of (weighted) polynomial least squares approximation then provides conditions on the number of samples required to achieve quasi-optimal approximation of f_L within a given space of polynomials V_L. However, this approach can be computationally expensive, as each evaluation of f_L requires the numerical solution of a PDE using a fine discretization. As an alternative, our proposed multilevel algorithm starts out with a least squares approximation of f_0 using a relatively large polynomial subspace V_0 and correspondingly many samples. To correct for the committed error $f - f_0$, the algorithm then adds polynomial approximations of $f_l - f_{l-1}$ that lie in subspaces V_l, $l \in \{1, \dots, L\}$.

Since we assume that $f_l \to f$ in an appropriate sense, the differences $f_l - f_{l-1}$ may be approximated using smaller polynomial subspaces for $l \to \infty$. Exploiting this fact, it is possible to obtain approximations with significantly reduced computational work. Indeed, we show that under certain conditions the work that the

multilevel method requires to attain an accuracy of $\epsilon > 0$ is the same as the work that regular least squares polynomial approximation would require if f could be evaluated exactly. It is clear that such a result is not always possible. For example, if f were constant, then polynomial least squares approximations in any fixed polynomial subspace would yield the exact solution given a sufficiently large sample size. This means that the work required to achieve an accuracy $\epsilon > 0$ would be bounded as $\epsilon \to 0$, which can clearly not be true for an algorithm that uses evaluations from approximate functions f_l that become more expensive to evaluate as $l \to \infty$. Instead, the computational work required for an accuracy of $\epsilon > 0$ in this case is determined by the convergence of $f_l \to f$ and by the work that is required for evaluations of f_l. Theorem 4 show that under certain conditions the two cases described above are dichotomous: the computational work of the multilevel method is either that of solving a single PDE or that of performing polynomial regression of a function that allows exact evaluations.

The remainder of this work is structured as follows. In Section 2, we review the theoretical analysis of weighted least squares approximation. In Section 3, we discuss different sampling strategies. We propose algorithms to sample the optimal distribution and we discuss the consequences of using perturbed distributions. In Section 4, we introduce a novel multilevel algorithm and prove our main results concerning the work and convergence of this algorithm. For situations in which the regularity of f and the convergence of f_l are not known, we propose an adaptive algorithm in Section 5. We discuss the applicability of our method to problems in uncertainty quantification in Section 6. Finally, we present numerical experiments in Section 7.

2. Weighted least squares polynomial approximation

In this section, we provide a short summary of the theory of weighted discrete least squares polynomial approximation, closely following [6]. Assume that we want to approximate a function $f \in L^2_\mu(\Gamma)$, where $\Gamma \subset \mathbb{R}^d$ and μ is a probability measure on Γ. The strategy of weighted discrete least squares polynomial approximation is to

- choose a finite-dimensional space $V \subset L^2_\mu(\Gamma)$ of polynomials on Γ
- choose a function $\rho\colon \Gamma \to \mathbb{R}$ that satisfies $\int_\Gamma \rho(\boldsymbol{y})\mu(\mathrm{d}\boldsymbol{y}) = 1$ and $\rho > 0$
- generate $N > 0$ independent random samples from the *sampling distribution* ν defined by $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} := \rho$,

$$\boldsymbol{y}_j \sim \nu, \quad j \in \{1, \dots, N\}.$$

  Here, $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ denotes the density, or Radon–Nikodym derivative, of the probability measure ν with respect to the reference measure μ.
- evaluate f at $\boldsymbol{y}_j$, $j \in \{1, \dots, N\}$
- define the *weight function* $w := \frac{1}{\rho}\colon \Gamma \to \mathbb{R}$
- and finally define the *weighted discrete least squares approximation*

$$\Pi_V f := \operatorname*{arg\,min}_{v \in V} \|f - v\|_N, \tag{2.1}$$

where

$$\|g\|_N^2 := \langle g, g\rangle_N := \frac{1}{N}\sum_{j=1}^N w(\boldsymbol{y}_j)|g(\boldsymbol{y}_j)|^2 \quad \forall g\colon \Gamma \to \mathbb{R}. \tag{2.2}$$

It is straightforward to show that the coefficients $\mathbf{v}$ of $\Pi_V f$ with respect to any basis $(B_j)_{j=1}^m$ of V are given by

$$\mathbf{G}\mathbf{v} = \mathbf{c}, \tag{2.3}$$

with $\mathbf{G}_{ij} := \langle B_i, B_j\rangle_N$, and $c_j := \langle f, B_j\rangle_N$, $i, j \in \{1, \dots, m\}$, assuming that $\mathbf{G}$ is invertible. If $\mathbf{G}$ is not invertible, then (2.1) has multiple solutions and we define $\Pi_V f$ as the one with the minimal $L^2_\mu(\Gamma)$ norm.

Remark 2.1. Assembling the matrix $\mathbf{G}$ requires $\mathcal{O}(m^2N)$ operations. However, using the fact that $\mathbf{G} = \mathbf{M}^\top\mathbf{M}$ for $\mathbf{M}_{ij} := N^{-1/2}\sqrt{w(\boldsymbol{y}_i)}B_j(\boldsymbol{y}_i)$, matrix vector products with $\mathbf{G}$ can be computed at the lower cost $\mathcal{O}(mN)$ as $\mathbf{G}\mathbf{x} = \mathbf{M}^\top(\mathbf{M}\mathbf{x})$. See Remark 4.5 below for the computation of products of the form $\mathbf{G}^{-1}\mathbf{x}$.

Since $w\rho = 1$, the semi-norm defined in (2.2) is a Monte Carlo approximation of the $L^2_\mu(\Gamma)$ norm. Therefore, we may expect that the error $\|f - \Pi_V f\|_{L^2_\mu(\Gamma)}$ is close to the optimal one,

$$e_{V,2}(f) := \min_{v\in V} \|f - v\|_{L^2_\mu(\Gamma)}. \tag{2.4}$$

Part (iii) of Theorem 2.2 below shows that this is true in expectation, provided that the number of samples N is coupled appropriately to the dimension $m = \dim V$ of the approximating polynomial subspace and provided that we ignore outcomes where $\mathbf{G}$ is ill-conditioned. For results in probability, we need to replace the best $L^2_\mu(\Gamma)$ approximation by the best approximation in a weighted supremum norm,

$$e_{V,w,\infty}(f) := \inf_{v\in V} \sup_{\boldsymbol{y}\in\Gamma} |f(\boldsymbol{y}) - v(\boldsymbol{y})|\sqrt{w(\boldsymbol{y})}. \tag{2.5}$$

Theorem 2.2 (Convergence of weighted least squares, [6], Thm. 2)**.** *For arbitrary $r > 0$, define*

$$\kappa := \frac{1/2 - 1/2\log 2}{1 + r}.$$

Assume that for all $\boldsymbol{y} \in \Gamma$ there exists $v \in V$ such that $v(\boldsymbol{y}) \neq 0$ and denote by $(B_j)_{j=1}^m$ an L^2_μ-orthonormal basis of V. Finally, assume that

$$K_{V,w} := \left\| w \sum_{j=1}^m B_j^2 \right\|_{L^\infty(\Gamma)} \leq \kappa \frac{N}{\log N}. \tag{2.6}$$

(i) *With probability larger than $1 - 2N^{-r}$, we have*

$$\|\mathbf{G} - \mathbf{I}\| \leq \frac{1}{2}, \tag{2.7}$$

where $\mathbf{G}$ is the matrix from (2.3), $\mathbf{I}$ is the identity matrix, and $\|\cdot\|$ denotes the spectral matrix norm.

(ii) *If $\|\mathbf{G} - \mathbf{I}\| \leq 1/2$, then for all f with $\sup_{\boldsymbol{y}\in\Gamma} |f(\boldsymbol{y})|\sqrt{w(\boldsymbol{y})} < \infty$, we have*

$$\|f - \Pi_V f\|_{L^2_\mu(\Gamma)} \leq (1 + \sqrt{2}) e_{V,w,\infty}(f).$$

(iii) *If $f \in L^2_\mu(\Gamma)$, then*

$$\mathbb{E}\,\|f - \Pi_V^{\mathrm{c}} f\|^2_{L^2_\mu(\Gamma)} \leq \left(1 + \frac{4\kappa}{\log N}\right) e^2_{V,2}(f) + 2\,\|f\|^2_{L^2_\mu(\Gamma)}\, N^{-r},$$

where $\mathbb{E}$ denotes the expectation with respect to the N-fold draw from the sampling distribution ν and

$$\Pi_V^{\mathrm{c}} f := \begin{cases} \Pi_V f & \text{if} \quad \|\mathbf{G} - \mathbf{I}\| \leq \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. It is proved in Theorem 2 of [6] that the bound in part (ii) holds for a fixed f with probability larger than $1 - 2N^{-r}$. A look at the proof reveals that the bound only depends on the event $\|\mathbf{G} - \mathbf{I}\| \leq 1/2$ and not on the specific choice of f. The remaining claims are exactly as in [6]. □

3. Sampling strategies

It was observed in [6] that the constant $K_{V,w}$ in (2.6) satisfies

$$\begin{aligned} m &= \int \sum_{j=1}^{m} |B_j(\boldsymbol{y})|^2 \mu(\mathrm{d}\boldsymbol{y}) \\ &\leq \left(\int w^{-1}(\boldsymbol{y}) \mu(\mathrm{d}\boldsymbol{y}) \right) \left\| w \sum_{j=1}^{m} B_j^2 \right\|_{L^\infty(\Gamma)} \\ &= K_{V,w} \end{aligned} \tag{3.1}$$

and that the inequality becomes an equality for the weight $w_V^* = {\rho_V^*}^{-1}$ that is associated with the density

$$\rho_V^*(\boldsymbol{y}) := \frac{1}{m} \sum_{j=1}^{m} |B_j(\boldsymbol{y})|^2. \tag{3.2}$$

For this choice, Theorem 2.2 roughly asserts that the number of samples required to determine a near-optimal approximation of f in an m-dimensional space V is smaller than $Cm \log m$ for some $C > 0$. In the remainder of this work, we refer to w_V^*, ρ_V^*, and

$$\nu_V^*: \quad \frac{\mathrm{d}\nu_V^*}{\mathrm{d}\mu} := \rho_V^* \tag{3.3}$$

as the *optimal* weight, density, and distribution, respectively. Since the optimal distribution ν_V^* depends on V, practical implementations need to address the question how to obtain samples from ν_V^* for general subspaces V. Furthermore, since ρ_V^* depends on V, the weight in $e_{V,w,\infty}(f)$ in part (ii) of Theorem 2.2 does as well. To address these issues, we present two types of results in this section.

First, we discuss how to obtain samples from ν_V^*. For the case where is a product domain equipped with a product measure, we propose a method for the generation of N samples whose computational work is bounded in expectation by the product KdN with a constant K that depends only on the measures μ_j. For non-product domains or measures, we briefly discuss how to use Markov chain Monte Carlo (MCMC) sampling for the generation of samples from approximate distributions and how perturbations of the sampling distributions affect the error estimates.

Second, we prove that the density of the optimal distribution ν_V^* associated with any downward closed polynomial subspace on $[0,1]^d$ with respect to the Lebesgue measure $\mathrm{d}\lambda$ satisfies

$$C^{-d} < \frac{\mathrm{d}\nu_V^*}{\mathrm{d}\lambda} \leq C^d p_d^\infty, \tag{3.4}$$

where $0 < C < \infty$ is independent of V, and p_d^∞ is the Lebesgue density of the d-dimensional arcsine distribution,

$$p_d^\infty(\boldsymbol{y}) := \prod_{j=1}^{d} \frac{1}{\pi \sqrt{y_j(1-y_j)}}. \tag{3.5}$$

The lower bound in (3.4) implies that the optimal weight w_V^* is bounded above by C^d, which can be used to make the error estimate in part (ii) of Theorem 2.2 more explicit. By the upper bound, we may use samples from the d-dimensional arcsine distribution instead of the optimal distribution. Indeed, the upper bound implies that the weight function w associated with the arcsine distribution satisfies $K_{V,w} \leq C^d m$. Thus, the required number of samples is increased at most by the factor C^d, which is independent of V. Preliminary numerical experiments indicate that the true factor is smaller than 4 even for $d = 10$. The advantages are that samples from the arcsine distribution can be generated efficiently, that we can use samples from the same distribution for all polynomial subspaces, and that the weight w is easy to analyze and independent of V.

3.1. Sampling from the optimal distribution

We now describe an efficient algorithm to obtain samples from ν_V^* in the case when Γ is a Cartesian product, μ is a product measure, and V is downward closed.

Definition 3.1 (Downward closedness)**.** Let $\mathbb{N} := \{0, 1, \ldots\}$. A set $\mathcal{I} \subset \mathbb{N}^d$ is called *downward closed* if $\eta \in \mathcal{I}$ implies $\eta' \in \mathcal{I}$ for any $\eta' \in \mathbb{N}^d$ with $\eta' \leq \eta$ componentwise.

A space V of polynomials on a Cartesian product domain $\Gamma = \prod_{j=1}^d I_j$ with $I_j \subset \mathbb{R}$ is called *downward closed* if it is the span of monomials,

$$V = \operatorname{span}\left\{ \boldsymbol{y}^\eta = \prod_{j=1}^d y_j^{\eta_j} : \eta \in \mathcal{I} \right\},$$

for some downward closed set $\mathcal{I} \subset \mathbb{N}^d$.

Remark 3.2. Observe that any non-trivial downward closed polynomial space V includes the constant functions and thus satisfies the assumption of Theorem 2.2 that for all $\boldsymbol{y} \in \Gamma$ there exists $v \in V$ with $v(\boldsymbol{y}) \neq 0$.

We first discuss the case $\Gamma = [0,1]^d$ and $\mu = \lambda$ the Lebesgue measure. For any downward closed subspace

$$V = \operatorname{span}\{\boldsymbol{y}^\eta : \eta \in \mathcal{I}\} \subset L^2_\lambda([0,1]^d)$$

with $\mathcal{I} \subset \mathbb{N}^d$ and $|\mathcal{I}| = \dim V = m$, an orthonormal basis is then given by

$$(P_\eta)_{\eta \in \mathcal{I}}$$

where

$$P_\eta(\boldsymbol{y}) := \prod_{j=1}^d P_{\eta_j}(y_j)$$

and $(P_n)_{n\in\mathbb{N}}$ are the Legendre polynomials on $[0,1]$, which are orthonormal with respect to the one-dimensional Lebesgue measure. By orthonormality, each P_η^2 may be interpreted as a probability density with respect to the Lebesgue measure. Thus,

$$\frac{\mathrm{d}\nu_V^*}{\mathrm{d}\lambda} = \rho_V^* = \frac{1}{m} \sum_{\eta \in \mathcal{I}} P_\eta^2$$

may be interpreted as mixture of m probability densities. An efficient strategy to obtain samples from ν_V^* is therefore to first choose $\eta \in \mathcal{I}$ at random and then generate a sample from the distribution with Lebesgue density P_η^2. Since $P_\eta^2 = \prod_{j=1}^d L_{\eta_j}^2$, samples from this distribution can be generated componentwise. Finally, to obtain samples from the univariate distributions with Lebesgue densities P_n^2, $n \in \mathbb{N}$, we use a rejection sampling method with the arcsine proposal density p_1^∞. By Theorem 1 of [30] the Legendre polynomials satisfy

$$|P_n(y)|^2 \leq 4e p_1^\infty(y) \quad \forall y \in [0,1]\ \forall n \in \mathbb{N}. \tag{3.6}$$

Therefore, the theory of rejection sampling ([12], Chap. 4.5) ensures that if we repeatedly generate $y \sim p_1^\infty$ and $U \sim \mathrm{Unif}(0,1)$ until $U \leq |P_n(y)|^2/(4ep_1^\infty(y))$ holds, then the resulting sample is exactly distributed according to P_n^2 and the required number of iterations until acceptance has a geometric distribution with mean $4e$. The total expected computational work for the generation of N samples from ν_V^* is thus $4eNd$, if we assume that the computation of $P_n^2(y)$ is $O(1)$. In practice, a 3-term recurrence formula whose work is bounded by $3n$ can be used to compute $P_n(y)$. This increases the upper bound for the expected work to $12eN\frac{1}{m}\sum_{\eta\in\mathcal{I}} |\eta|_1$.

Equation (3.6) holds more generally for probability measures on $[0,1]$ with Lebesgue densities of the form $\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} = C(\alpha,\beta) y^\alpha (1-y)^\beta$, $\alpha, \beta \geq -1/2$ ([30], Thm. 1). The bound on the associated orthogonal polynomials $(P_n^{\alpha,\beta})_{n\in\mathbb{N}}$, which are commonly called Jacobi polynomials, is

$$|P_n^{\alpha,\beta}(y)|^2 \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \leq 2e(2 + \sqrt{\alpha^2+\beta^2}) p_1^\infty(y) \qquad \forall y \in [0,1] \quad \forall n \in \mathbb{N}.$$

Even more generally, the same inequality holds with a constant C_μ independent of $\boldsymbol{y}$ and n for orthogonal polynomials with respect to a wide class of measures μ that are absolutely continuous with respect to the Lebesgue measure on $[0,1]$ ([33], Thm. 12.1.4). When C_μ is unknown, however, rejection sampling cannot be applied. As a substitute, we could use MCMC sampling (which we also discuss below as an alternative method to sample directly from ν^* in cases when no product structure of Γ or μ can be exploited). The error due to the fact that the resulting samples would not be distributed exactly according to $|P_n|^2$ can be controlled using Proposition 3.3 below.

For orthonormal polynomials $(H_n)_{n\in\mathbb{N}}$ with respect to rapidly decaying measures supported on the whole real line, such as Gaussian measures, it is shown in [24] that $|H_n(y)|^2 \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ is exponentially concentrated in an interval $[-a_n, a_n]$ with $C^{-1}n^b \le a_n \le Cn^b$ for some $b > 0$ and $C > 0$ depending on μ, and that for some C_μ

$$|H_n(y)|^2 \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \le C_\mu \frac{a_n}{4} \left|1 - \frac{y}{a_n}\right|^{-1/2} \qquad \forall y \in [-a_n, a_n] \quad \forall n \in \mathbb{N}.$$

Together with the stability result in Proposition 3.3 below, this shows that the previous results can be transfered to measures on the real line, if we simply ignore the mass outside $[-a_n, a_n]$ and apply rejection sampling or Markov chain methods with the proposal density $\frac{a_n}{4}|1 - \frac{y}{a_n}|^{-1/2}$. Alternatively, a different result in [24] shows that on $[-a_n, a_n]$ the density $|H_n(y)|^2 \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ is bounded by the uniform probability density up to a factor that grows sublinearly in the polynomial degree n.

The previous example motivates looking at situations where exact sampling from the optimal distribution is not practical or feasible, and one resorts to inexact sampling, instead. This will also be the case when Markov chain Monte Carlo samplers are used, as discussed below. The next proposition quantifies the effect of inexact sampling in the results of Theorem 2.2.

Proposition 3.3 (Stability with respect to perturbations of the sampling density)**.** All results in Theorem 2.2 that are valid for the optimal choice ν_V^* with $\frac{\mathrm{d}\nu_V^*}{\mathrm{d}\mu} = \rho_V^*$ of the sampling distribution hold true if we instead use samples from a distribution $\tilde{\nu}$ (but keep the weight function $w_V^* = 1/\rho_V^*$) that satisfies

$$\|\tilde{\rho}/\rho_V^* - 1\|_{L^\infty} \le c$$

or

$$\|\tilde{\nu} - \nu_V^*\|_{\mathrm{TV}} := \frac{1}{2}\|\tilde{\rho} - \rho_V^*\|_{L^1_\mu(\Gamma)} \le \frac{c}{2m},$$

for $\tilde{\rho} := \frac{\mathrm{d}\tilde{\nu}}{\mathrm{d}\mu}$ and $c \in [0, 1/2)$, provided that we replace κ by $\frac{(1-2c)^4}{(1+r)10}$.

Proof. The proof of Theorem 2.2 in [6] is based on large deviation bounds for the matrix $\mathbf{G}$ of (2.3). In particular, it is based on the observation that $\mathbf{G}$ is a Monte Carlo average,

$$\mathbf{G} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i,$$

of independent and identically distributed matrices

$$\mathbf{X}_i := (w_V^*(\boldsymbol{y}_i)B_j(\boldsymbol{y}_i)B_k(\boldsymbol{y}_i))_{j,k\in\{1,\dots,m\}} \quad \text{with } \boldsymbol{y}_i \sim \nu_V^*$$

that satisfy

$$\mathbb{E}\mathbf{X}_i = \mathbf{I}$$

by $L^2_\mu(\Gamma)$-orthonormality of the basis polynomials B_j, $j \in \{1,\dots,m\}$ and $\|\mathbf{X}_i\| \le m$ almost surely by definition of w_V^*. A Chernoff inequality for matrices then provides the bound on $\mathbb{P}(\|\mathbf{G} - \mathbf{I}\| \le 1/2)$ in part

(i) of Theorem 2.2 from which everything else follows. The crucial insight is that this inequality permits small perturbations of the expected value. Indeed, if we replace ν_V^* by $\tilde{\nu}$ in the definition of $\mathbf{X}_i$, then Theorem 1.1 of [34] yields the same bound on $\mathbb{P}(\|\mathbf{G}-\mathbf{I}\| \leq 1/2)$, with the new value of κ, provided that $\|\boldsymbol{M}\| = \sup_{\|z\|=1}\langle \boldsymbol{M}z, z\rangle \leq c$ for $\boldsymbol{M} := \mathbb{E}\mathbf{X}_i - \mathbf{I}$ (note that $\mu_{\max}/R$ from Thm. 1.1 of [34] is then larger than $(1-c)N/m \geq \log(m)(1+r)10/(1-2c)^3$ and that $(1+\delta)^{-(1+\delta)}\exp(\delta)$ from the same theorem is smaller than $\exp(-(1-2c)^3/10)$ for $(1+\delta) := 3/(2(1+c))$). To show $\|\boldsymbol{M}\| \leq c$, we observe that the entries of $\mathbb{E}X_i$ are given by $\int_\Gamma w_V^* B_j B_k \,\mathrm{d}\tilde{\nu} = \int_\Gamma \tilde{\rho}/\rho_V^* B_j B_k \,\mathrm{d}\mu$. Hence, we obtain the representation

$$\langle \boldsymbol{M}z, z\rangle = \int_\Gamma (\tilde{\rho}/\rho_V^* - 1)\mathrm{d}\pi_z,$$

where π_z is the probability measure defined by $\frac{\mathrm{d}\pi_z}{\mathrm{d}\mu} = (\sum_{j=1}^m z_j B_j)^2$. This shows that $\|\boldsymbol{M}\| \leq c$ if $\|\tilde{\rho}/\rho_V^* - 1\|_{L^\infty} \leq c$. Furthermore, since

$$\begin{aligned}\frac{\mathrm{d}\pi_z}{\mathrm{d}\nu_V^*} &= \frac{\mathrm{d}\pi_z}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\nu_V^*}\\ &= \left(\sum_{j=1}^m z_j B_j\right)^2 \frac{m}{\sum_{j=1}^m B_j^2}\\ &\leq m\end{aligned}$$

by the Cauchy–Schwarz inequality, the same estimate holds if $\|\tilde{\rho}/\rho_V^* - 1\|_{L^1_{\nu_V^*}(\Gamma)} = \|\tilde{\rho} - \rho_V^*\|_{L^1_\mu(\Gamma)} \leq c/m$. □

So far, we have assumed that Γ and μ exhibit product structure, which allowed us to generate samples coordinate-wise, exploiting known bounds on univariate orthogonal polynomials. For more general cases, we now briefly discuss Metropolized independent sampling, which is a simple MCMC algorithm, for the generation of samples from the optimal distribution ν_V^*. For an extensive treatment of the theory of MCMC algorithms we refer to [26].

The general strategy of MCMC algorithms for the generation of samples from ν_V^* is to construct a Markov chain for which ν_V^* is an invariant distribution. Ergodic theory then shows that under some assumptions the location of this Markov chain after $n \gg 1$ steps is approximately distributed according to ν_V^*. Metropolis–Hastings algorithms are MCMC algorithms that construct Markov chains based on user-specified *proposal densities* $p(\boldsymbol{y}, \cdot)$, $\boldsymbol{y} \in \Gamma$ (with respect to μ) and a rejection step to ensure convergence to the desired limit distribution ν_V^*. More specifically, the transition kernel of a Metropolis–Hastings algorithm has the form

$$K(\boldsymbol{y}, \mathrm{d}\boldsymbol{y}') := \begin{cases} p(\boldsymbol{y}, \boldsymbol{y}')\min\left\{1, \frac{\rho_V^*(\boldsymbol{y}')p(\boldsymbol{y}', \boldsymbol{y})}{\rho_V^*(\boldsymbol{y})p(\boldsymbol{y}, \boldsymbol{y}')}\right\}\mu(\mathrm{d}\boldsymbol{y}') & \text{if} \quad \boldsymbol{y}' \neq \boldsymbol{y} \\ 1 - \int_{z\neq\boldsymbol{y}} p(\boldsymbol{y}, z)\min\left\{1, \frac{\rho_V^*(z)p(z, \boldsymbol{y})}{\rho_V^*(\boldsymbol{y})p(\boldsymbol{y}, z)}\right\}\mu(\mathrm{d}z) & \text{if} \quad \boldsymbol{y}' = \boldsymbol{y}. \end{cases} \tag{3.7}$$

This kernel can be interpreted (and implemented) as proposing a transition from the current state $\boldsymbol{y}$ to a new state $\boldsymbol{y}'$ drawn from the density $p(\boldsymbol{y}, \cdot)$, and rejecting this transition with a certain probability determined by the values of ρ_V^* and p at the current state $\boldsymbol{y}$ and the proposed state $\boldsymbol{y}'$. The rejection probability is designed to ensure the detailed balance condition $\nu_V^*(\mathrm{d}\boldsymbol{y})K(\boldsymbol{y}, \mathrm{d}\boldsymbol{y}') = \nu_V^*(\mathrm{d}\boldsymbol{y}')K(\boldsymbol{y}', \mathrm{d}\boldsymbol{y})$, which in turn guarantees that ν_V^* is invariant under K.

Metropolized independent sampling is the name of the subset of Metropolis–Hastings algorithms for which the proposal density p is independent of the current state $\boldsymbol{y}$. If we denote the corresponding state-independent proposal density by $p(\boldsymbol{y}')$ and define $g := \inf_{\boldsymbol{y}\in\Gamma} p(\boldsymbol{y})/\rho_V^*(\boldsymbol{y})$, then it can be shown Section 3.2.2 of [25] that starting from any distribution π we have the bound

$$\|K^n\pi - \nu_V^*\|_{\mathrm{TV}} \leq 2(1-g)^n \tag{3.8}$$

for the total variation distance between the nth step probability distribution $K^n\pi$ of the Markov chain and the target distribution ν_V^*. This means that if the proposal density satisfies $g := \inf_{\boldsymbol{y}\in\Gamma} p(\boldsymbol{y})/\rho_V^*(\boldsymbol{y}) > 0$, then $n := g^{-1}\log(24m)$ Markov chain steps suffice to ensure that

$$\|K^n\pi - \nu_V^*\|_{\mathrm{TV}} \le 2(1-g)^{g^{-1}\log(24m)} \le \frac{1}{12m},$$

as required by Proposition 3.3. To generate $N > 0$ independent samples from $K^n\pi$, we have to run N independent copies of the Markov chain, which differs from the more common practice to use N successive, thus dependent, steps of a single Markov chain.

3.2. Sampling from the arcsine distribution

In Proposition 3.4 below, we determine lower and upper bounds for the optimal sampling distributions of downward closed polynomial subspaces on $[0,1]^d$. Although we restrict ourselves to the Lebesgue measure, the results can be extended verbatim to more general measures on the hypercube.

The lower bound can be used to make the bound in Theorem 2.2 more precise. Indeed, it implies that the weight $w_V^* = \frac{\mathrm{d}\lambda}{\mathrm{d}\nu_V^*}$ appearing in $e_{V,w_V^*,\infty}$ satisfies

$$w_V^* \le C^d \tag{3.9}$$

The upper bound provides an alternative sampling strategy: Instead of sampling from the optimal distribution, we may simply sample from the arcsine distribution with Lebesgue density p_d^∞ without using Acceptance/Rejection or Markov chain methods. Indeed, using the arcsine distribution for sample generation amounts to using the sampling density $\rho = p_d^\infty$ and the weight function $w := (p_d^\infty)^{-1}$ in Section 2. Hence, the upper bound shows that the corresponding constant $K_{V,w}$ in Theorem 2.2 satisfies

$$\begin{aligned} K_{V,w} &= \left\| w \sum_{j=1}^m P_j^2 \right\|_{L^\infty(\Gamma)} \\ &= \left\| wm \frac{\mathrm{d}\nu_V^*}{\mathrm{d}\lambda} \right\|_{L^\infty(\Gamma)} \\ &\le C^d m, \end{aligned} \tag{3.10}$$

which is larger than the optimal value, m, only by the factor C^d. The advantages are that exact and independent samples from the univariate arcsine distribution can be generated efficiently as $(\sin(X)+1)/2$ for a uniform random variable X on $[-\pi/2,\pi/2]$, that we can use samples from the same distribution for all polynomial subspaces, and that the weight w that enters the error estimate in Theorem 2.2 through $e_{V,w,\infty}$ is known explicitly, vanishes at the boundary, and is independent of V.

Proposition 3.4 (Bounds on the optimal distribution)**.** There exists a constant $0 < C < \infty$ such that the optimal sampling distribution ν_V^* associated with any finite-dimensional downward closed space V of polynomials on $[0,1]^d$ equipped with the Lebesgue measure satisfies

$$C^{-d} \le \frac{\mathrm{d}\nu_V^*}{\mathrm{d}\lambda} \le C^d p_d^\infty. \tag{3.11}$$

Proof. Equation (3.11) was shown to hold for the univariate optimal sampling distributions ν_k^* associated with univariate spaces of polynomials of degree less than or equal to $k \in \mathbb{N}$ on $[0,1]$ in equation (7.14) of [27]. We prove the case $d > 1$ by induction.

Since V is downward closed, we have

$$V = \mathrm{span}_{\eta\in\mathcal{I}\subset\mathbb{N}^d}\{P_\eta(\boldsymbol{y}) := P_{\eta_1}(y_1)\cdots P_{\eta_d}(y_d)\}$$

for some multi-index set $\mathcal{I} \subset \mathbb{N}^d$. We define the sliced multi-index sets $\mathcal{I}_j := \{\eta \in \mathcal{I} : \eta_1 = j\}$, $j \in \mathbb{N}$ and the corresponding spaces

$$\tilde{V}_j := \operatorname{span}_{\eta \in \mathcal{I}_j \subset \mathbb{N}^d} \{\tilde{P}_\eta(\tilde{\boldsymbol{y}}) := P_{\eta_2}(y_2) \cdots P_{\eta_d}(y_d)\}$$

of polynomials on $[0,1]^{d-1}$ with associated optimal distributions $\tilde{\nu}_j$ on $[0,1]^{d-1}$. This allows us to write

$$\begin{aligned}\frac{\mathrm{d}\nu_V^*}{\mathrm{d}\lambda}(\boldsymbol{y}) &= \rho_V^*(\boldsymbol{y}) \\ &= \frac{1}{|\mathcal{I}|} \sum_{\eta \in \mathcal{I}} P_\eta^2(\boldsymbol{y}) \\ &= \frac{1}{|\mathcal{I}|} \sum_{j \in \mathbb{N}} P_j^2(y_1) \sum_{\eta \in \mathcal{I}_j} \tilde{P}_\eta^2(\tilde{\boldsymbol{y}}) \\ &= \frac{1}{|\mathcal{I}|} \sum_{j \in \mathbb{N}} P_j^2(y_1) |\mathcal{I}_j| \frac{\mathrm{d}\tilde{\nu}_j}{\mathrm{d}\lambda}(\tilde{\boldsymbol{y}}),\end{aligned}$$

which, by the induction hypothesis for the case $d-1$, entails

$$C^{-(d-1)} A(y_1) \leq \frac{\mathrm{d}\nu_V^*}{\mathrm{d}\lambda}(\boldsymbol{y}) \leq A(y_1) C^{d-1} p_{d-1}^\infty(\tilde{\boldsymbol{y}}) \tag{3.12}$$

with

$$A(y_1) := \frac{1}{|\mathcal{I}|} \sum_{j \in \mathbb{N}} P_j^2(y_1) |\mathcal{I}_j|.$$

We now use the fact that $A(y_1)$ can be written as a weighted average of the univariate densities $\frac{\mathrm{d}\nu_k^*}{\mathrm{d}\lambda}(y_1) = \frac{1}{k} \sum_{j=1}^k P_j^2(y_1)$:

$$A(y_1) = \frac{1}{\sum_{k=1}^\infty p_k} \sum_{k=1}^\infty p_k \frac{\mathrm{d}\nu_{p_k}^*}{\mathrm{d}\lambda}(y_1)$$

with $p_k := |\{j : |\mathcal{I}_j| \geq k\}|$.

Together with the induction hypothesis for the case $d = 1$, this implies

$$C^{-1} \leq A(y_1) \leq C p_1^\infty(y_1),$$

which, when inserted into (3.12), yields

$$C^{-d} = C^{-(d-1)} C^{-1} \leq \frac{\mathrm{d}\nu_V^*}{\mathrm{d}\lambda}(\boldsymbol{y}) \leq C^{d-1} C p_1^\infty(\boldsymbol{y}_1) p_{d-1}^\infty(\tilde{\boldsymbol{y}}) = C^d p_d^\infty(\boldsymbol{y}).$$

□

4. Multilevel weighted least squares approximation

In this section, we define a multilevel weighted polynomial least squares method and establish convergence rates for the approximation of a function $f_\infty \colon \Gamma \subset \mathbb{R}^d \to \mathbb{R}$, $d \in \mathbb{N} \cup \{\infty\}$ in a normed vector space $(F, \|\cdot\|_F) \hookrightarrow (L^2_\mu(\Gamma), \|\cdot\|_{L^2_\mu(\Gamma)})$ of continuous functions on Γ, under the following assumptions.

– **A1** (Convergence of approximations)**.** There exist functions $f_n \in F$, $n \geq 1$ such that

$$\begin{aligned}\|f_\infty - f_n\|_F &\lesssim n^{-\beta_s} \\ \|f_\infty - f_n\|_{L^2_\mu(\Gamma)} &\lesssim n^{-\beta_w}\end{aligned}$$

for some $\beta_s > 0$ and $\beta_w \geq \beta_s$.

– **A2(p)** (Polynomial approximability)**.** There exist downward closed spaces of polynomials V_m, $m \geq 1$ on Γ such that

$$\dim V_m \lesssim m^{\sigma},$$
$$e_{m,p}(F) \lesssim m^{-\alpha}$$

for some $\sigma > 0, \alpha > 0$, and $p = 2$ or $p = \infty$, where $e_{m,2}(F) := \sup_{f \in F} \frac{e_{V_m,2}(f)}{\|f\|_F}$ and $e_{m,\infty}(F) := \sup_{f \in F} \frac{e_{V_m,w_m^*,\infty}(f)}{\|f\|_F}$.

– **A3** (Sample work)**.** The work required for a single evaluation of f_n satisfies $\text{Work}(f_n) \lesssim n^{\gamma}$ for some $\gamma > 0$.

We use $\lesssim$ to denote inequalities that hold up to factors that are independent of n and m.

Remark 4.1. In Assumption A2(p), we have introduced the exponent σ, which in contrast to previous sections may be different from 1, to be able to apply our results with common sequences of polynomial subspaces without the need for reparametrization.

Example 4.2 (Polynomial approximability)**.**

– For univariate Sobolev spaces $F = H^{\alpha}(\Gamma)$, $\Gamma = (0,1)$ with $\alpha > 0$, Theorem 1 in [32] shows that

$$e_{m,2}(H^{\alpha}(\Gamma)) \lesssim m^{-\alpha}$$

for the space V_m of univariate polynomials with degree less than m and for $\mu = \lambda$ the Lebesgue measure. Analogous results also hold in higher dimensions. Here, optimal sequences of polynomial approximation spaces depend on the available smoothness. In particular, optimal polynomial approximation spaces for functions in Sobolev spaces $H^{\alpha}(\Gamma)$ with $\Gamma \subset \mathbb{R}^d$ and $\alpha > 0$ are of total degree type, whereas functions in Sobolev spaces $H^{\alpha}_{\text{mix}}(\Gamma)$ of dominating mixed smoothness can be optimally approximated by hyperbolic cross polynomial spaces [11].
Similar results for the best approximation in the supremum norm hold for functions in Hölder spaces $F = C^{s,t}(\Gamma)$, $s \in \mathbb{N}$, $t \in [0,1]$ ([3], Thm. 2) (and their dominating mixed smoothness analogues).

– Alternatively, we may simply define the space F *via* polynomial approximability of its elements. Assume that we have a sequence $(V_m)_{m=1}^{\infty}$ of downward closed polynomial spaces on $\Gamma \subset \mathbb{R}^d$ with $d \in \mathbb{N} \cup \{\infty\}$. If for some $\alpha > 0$ we define

$$F := \left\{ f \colon \Gamma \to \mathbb{R} : \|f\|_F := \sup_{m \in \mathbb{N}} e_{V_m,p}(f) m^{\alpha} < \infty \right\}$$

with the auxiliary definition $V_0 := \{0\}$, then it is easy to show that $\|\cdot\|_F$ is a norm of F and that Assumption 2(p) holds with the given α. The choice of the sequence of subspaces V_m can be based on truncating a orthogonal decomposition of $L^2_{\mu}(\Gamma)$ such as to include only basis functions whose contribution is above a given threshold in V_m. For more information on this construction, see Section 5 and [10,17].

We now define the multilevel least squares method for a fixed number of levels $L \in \mathbb{N}$. We introduce the subsequences

$$m_k := M \exp(k/(\sigma + \alpha)), \quad k \in \{0, \dots, L\} \tag{4.1}$$

and

$$n_l := \exp(l/(\gamma + \beta_s)), \quad l \in \{0, \dots, L\}$$

with $M := \exp(L\delta)$, $\delta := \frac{\beta_w - \beta_s}{\alpha(\gamma + \beta_s)} \geq 0$ if $\gamma/\beta_s > \sigma/\alpha$ and $M := 1$ else. For our analysis we assume that m and n can take non-integer values; in practice, rounding up to the nearest integer increases the required work only by a constant factor. Abusing of notation, we keep the simple notation V_k, $e_{k,p}$, and f_l for the quantities V_{m_k}, $e_{m_k,p}$, and f_{n_l}, respectively.

Next, we draw independent, identically distributed, random samples

$$\Gamma_k = \{\boldsymbol{y}_{k,1}, \dots, \boldsymbol{y}_{k,|\Gamma_k|}\} \subset \Gamma, \quad k \in \{0, \dots, L\}$$

with $\boldsymbol{y}_{k,j} \sim \nu_k^*$, where $\nu_k^* := \nu_{V_k}^*$ is the optimal sampling distribution of V_k from (3.3). To ensure accuracy of our approximations, we couple the numbers of samples to the dimensions of the polynomial spaces *via*

$$m_k^\sigma \leq \kappa \frac{|\Gamma_k|}{\log |\Gamma_k|} \leq 2 m_k^\sigma \ \ \forall k \in \{0, \dots, L\}, \text{ where } \kappa := \frac{1 - \log 2}{2 + 2L}. \tag{4.2}$$

By (3.1), this guarantees that the assumption of Theorem 2.2 is satisfied with $r = L$. Alternatively, we may replace κ by $C^{-d}\kappa$ with C from Proposition 3.4 if Γ and $\mu \ll \lambda$ are products and if we use the arcsine distribution to generate samples, or we may choose κ as in Proposition 3.3 if we use samples that are only approximately distributed according to the optimal distribution.

Finally, we denote by $\Pi_k \colon F \to V_k$ the random weighted least squares approximation using evaluations in Γ_k, $k \in \{0, \dots, L\}$ and define the multilevel method

$$\begin{aligned} \mathcal{S}_L(f_\infty) &:= \Pi_L f_0 + \sum_{l=1}^{L} \Pi_{L-l}(f_l - f_{l-1}) \\ &= \sum_{l=0}^{L} \Pi_{L-l}(f_l - f_{l-1}) \end{aligned} \tag{4.3}$$

where we used the auxiliary definition $f_{-1} := 0$.

To clarify (4.3), let us summarize the common case where $f(\boldsymbol{y})$ is a scalar quantity of interest $Q(u_{\boldsymbol{y}})$ of the solution $u_{\boldsymbol{y}}$ to some PDE with parameters $\boldsymbol{y}$, and where $f_n(\boldsymbol{y})$ is the corresponding approximation $Q(u_{\boldsymbol{y},n})$ obtained by solving the PDE with a finite element solver of maximal element diameter $h := n^{-1}$. In this case, we start out by solving the PDE with a coarse resolution h_0 for a large number $|\Gamma_L|$ of randomly chosen values of the parameters $\boldsymbol{y}$ and extrapolating these results to the entire parameter domain by means of a weighted least squares approximation in a large polynomial subspace V_L. Next, to reduce the error due to the low resolution h_0, we compute the difference between using h_1 and h_0 for a smaller number of $|\Gamma_{L-1}|$ samples and extrapolate this difference to the entire parameter domain again by means of another weighted least squares approximation in a smaller space V_{L-1}. This process is continued until we arrive at the difference $f_L - f_{L-1}$, which is of smaller magnitude and can thus be extrapolated at roughly the same accuracy as the previous levels using only very few samples.

The computations in the proofs of below are similar to those appearing in multilevel Monte Carlo methods [14], though some more care has to be taken care about the choice of norms and about failure probabilities. We denote by $\lesssim$ any inequality that holds up to a factor depending only on $\alpha, \beta_s, \beta_w, \gamma$ and on the factors from assumptions A1, A2(p) and A3.

Theorem 4.3 (Convergence in probability)**.** *Denote by*

$$\text{Work}\left(\mathcal{S}_L(f_\infty)\right) := |\Gamma_L| \text{Work}\left(f_0\right) + \sum_{l=1}^{L} |\Gamma_{L-l}| \left(\text{Work}\left(f_l\right) + \text{Work}\left(f_{l-1}\right)\right) \tag{4.4}$$

the work that $\mathcal{S}_L(f_\infty)$ requires for evaluations of the functions f_l, $l \in \{0,\dots,L\}$. Define

$$\lambda := \begin{cases} \sigma/\alpha & \text{if} \quad \gamma/\beta_s \le \sigma/\alpha \\ \theta\gamma/\beta_s + (1-\theta)\sigma/\alpha \quad \text{with} \quad \theta := \beta_s/\beta_w & \text{if} \quad \gamma/\beta_s > \sigma/\alpha \end{cases}$$

and

$$t := \begin{cases} 2 & \text{if} \quad \gamma/\beta_s < \sigma/\alpha \\ 3+\sigma/\alpha & \text{if} \quad \gamma/\beta_s = \sigma/\alpha \\ 1 & \text{if} \quad \gamma/\beta_s > \sigma/\alpha \quad \text{and} \quad \beta_w = \beta_s \\ 2 & \text{if} \quad \gamma/\beta_s > \sigma/\alpha \quad \text{and} \quad \beta_w > \beta_s \end{cases}.$$

Let $0 < \epsilon \lesssim 1$. If Assumptions A1, A2(∞), and A3 hold, then we may choose $L \in \mathbb{N}$ such that

$$\text{Work}\,(\mathcal{S}_L(f_\infty)) \lesssim \epsilon^{-\lambda} |\log\epsilon|^t \log|\log\epsilon|,$$

and such that in an event E with $\mathbb{P}(E^c) \lesssim \epsilon^{\log|\log\epsilon|}$ the multilevel approximation satisfies

$$\|f_\infty - \mathcal{S}_L(f_\infty)\|_{L^2_\mu(\Gamma)} \le \epsilon. \tag{4.5}$$

Proof. The strategy of this proof is to establish bounds on $\text{Work}\,(\mathcal{S}_L(f_\infty))$ and $\|f_\infty - \mathcal{S}_L(f_\infty)\|_{L^2_\mu(\Gamma)}$ for arbitrary $L \in \mathbb{N}$ first, and then to show that, for the right choice of L, the latter is smaller than ϵ and the former is bounded by $\epsilon^{-\lambda}|\log\epsilon|^t \log|\log\epsilon|$.

Work bounds. We may deduce immediately from (4.2) the rough upper bound

$$\sqrt{|\Gamma_k|} \le \frac{|\Gamma_k|}{\log|\Gamma_k|} \le \frac{2}{\kappa} M^\sigma \exp\left(k\frac{\sigma}{\sigma+\alpha}\right) \lesssim (L+1)M^\sigma \exp\left(k\frac{\sigma}{\sigma+\alpha}\right)$$

on the number of samples at level $k \in \{0,\dots,L\}$. Using (4.2) again and inserting the previous estimate, we obtain the finer estimate

$$\begin{aligned} |\Gamma_k| &\le (L+1)M^\sigma \exp\left(k\frac{\sigma}{\sigma+\alpha}\right) \log|\Gamma_k| \\ &\lesssim (L+1)M^\sigma(\log(L+1)+\log M^\sigma)\exp\left(k\frac{\sigma}{\sigma+\alpha}\right)(k+1). \end{aligned}$$

Since

$$\text{Work}\,(f_l) + \text{Work}\,(f_{l-1}) \lesssim \exp\left(l\frac{\gamma}{\gamma+\beta_s}\right)$$

by Assumption A3, we may conclude that

$$\begin{aligned} \text{Work}\,(\mathcal{S}_L(f_\infty)) &\lesssim (L+1)M^\sigma(\log(L+1)+\log M^\sigma)\sum_{l=0}^{L}\exp\left((L-l)\frac{\sigma}{\sigma+\alpha}\right)(L-l+1)\exp\left(l\frac{\gamma}{\gamma+\beta_s}\right) \\ &= (L+1)M^\sigma(\log(L+1)+\log M^\sigma)\exp\left(L\frac{\sigma}{\sigma+\alpha}\right) \\ &\quad \times \sum_{l=0}^{L}\exp\left(-l\left(\frac{\sigma}{\sigma+\alpha}-\frac{\gamma}{\gamma+\beta_s}\right)\right)(L-l+1). \end{aligned} \tag{4.6}$$

We now distinguish three cases.

(a) $\gamma/\beta_s < \sigma/\alpha$: In this case $\sigma/(\sigma+\alpha) > \gamma/(\gamma+\beta_s)$. Thus, the sum on the right-hand side of (4.6) satisfies

$$\sum_{l=0}^{L} \exp\left(-l\left(\frac{\sigma}{\sigma+\alpha} - \frac{\gamma}{\gamma+\beta_s}\right)\right)(L-l+1) \lesssim (L+1)\sum_{l=0}^{L} \exp\left(-l\left(\frac{\sigma}{\sigma+\alpha} - \frac{\gamma}{\gamma+\beta_s}\right)\right) \lesssim L+1.$$

Together with the fact that $M = 1$ in the case under consideration, this shows that

$$\text{Work}\left(\mathcal{S}_L(f_\infty)\right) \lesssim \exp\left(L\frac{\sigma}{\sigma+\alpha}\right)(L+1)^2\log(L+1).$$

(b) $\gamma/\beta_s = \sigma/\alpha$: In this case $\sigma/(\sigma+\alpha) = \gamma/(\gamma+\beta_s)$. Thus, the sum on the right-hand side of (4.6) equals $\sum_{l=0}^{L}(L-l+1) \lesssim (L+1)^2$ and we obtain

$$\text{Work}\left(\mathcal{S}_L(f_\infty)\right) \lesssim \exp\left(L\frac{\sigma}{\sigma+\alpha}\right)(L+1)^3\log(L+1).$$

since $M = 1$.

(c) $\gamma/\beta_s > \sigma/\alpha$: In this case $\sigma/(\sigma+\alpha) < \gamma/(\gamma+\beta_s)$. Thus, the sum on the right-hand side of (4.6) satisfies

$$\begin{aligned}
&\sum_{l=0}^{L} \exp\left(-l\left(\frac{\sigma}{\sigma+\alpha} - \frac{\gamma}{\gamma+\beta_s}\right)\right)(L-l+1) \\
&\quad= \exp\left(L\left(\frac{\gamma}{\gamma+\beta_s} - \frac{\sigma}{\sigma+\alpha}\right)\right)\sum_{l=0}^{L} \exp\left(-l\left(\frac{\gamma}{\gamma+\beta_s} - \frac{\sigma}{\sigma+\alpha}\right)\right)(l+1) \\
&\quad\lesssim \exp\left(L\left(\frac{\gamma}{\gamma+\beta_s} - \frac{\sigma}{\sigma+\alpha}\right)\right).
\end{aligned}$$

If $\beta_w = \beta_s$, then $M = 1$ and we obtain

$$\begin{aligned}
\text{Work}\left(\mathcal{S}_L(f_\infty)\right) &\lesssim (L+1)M^\sigma(\log(L+1) + \log M^\sigma)\exp\left(L\frac{\gamma}{\gamma+\beta_s}\right) \\
&\lesssim \exp\left(L\left(\frac{\gamma}{\gamma+\beta_s}\right)\right)(L+1)\log(L+1).
\end{aligned}$$

If instead $\beta_w > \beta_s$, then $M = \exp(\delta L)$ and we obtain

$$\begin{aligned}
\text{Work}\left(\mathcal{S}_L(f_\infty)\right) &\lesssim (L+1)M^\sigma(\log(L+1) + \log M^\sigma)\exp\left(L\frac{\gamma}{\gamma+\beta_s}\right) \\
&\lesssim \exp\left(L\left(\frac{\gamma}{\gamma+\beta_s} + \sigma\delta\right)\right)(L+1)^2\log(L+1).
\end{aligned}$$

Residual bounds. First, we show that with high probability

$$\|\text{Id} - \Pi_k\|_{F\to L^2_\mu(\Gamma)} \lesssim M^{-\alpha}\exp\left(-k\alpha/(\sigma+\alpha)\right) \quad \forall k \in \{0,\dots,L\}. \tag{4.7}$$

By part (ii) of Theorem 2.2 together with Assumption A2(∞), it suffices to show that the event

$$E := \{\|\mathbf{G}_k - \mathbf{I}_k\| \le 1/2 \ \forall k \in \mathbb{N}\}$$

has a high probability, where $\mathbf{G}_k$ is the Gramian matrix from (2.3). But by the first part of the same theorem, the complementary probability that $\|\mathbf{G}_k - \mathbf{I}_k\| \le 1/2$ for a fixed $k \in \mathbb{N}$ decays as the number of samples $|\Gamma_k|$

increases. Since the sets Γ_k grow exponentially in k, by (4.2), we may conclude using a crude zeroth moment estimate and a geometric series bound:

$$\begin{aligned}\mathbb{P}(E^c) &= \mathbb{P}\left(\exists k \in \mathbb{N} : \|\mathbf{G}_k - \mathbf{I}_k\| > 1/2\right)\\ &\leq \sum_{k=0}^{\infty} \mathbb{P}(\|\mathbf{G}_k - \mathbf{I}_k\| > 1/2)\\ &\leq 2\sum_{k=0}^{\infty} |\Gamma_k|^{-L}\\ &\leq 2\kappa^L M^{-\sigma L} \sum_{k=0}^{\infty} \exp\left(-kL\frac{\sigma}{\sigma+\alpha}\right)\\ &= \frac{2\kappa^L M^{-\sigma L}}{1-\exp\left(-L\frac{\sigma}{\sigma+\alpha}\right)}\\ &\lesssim L^{-L}.\end{aligned} \tag{4.8}$$

Assuming now that the samples Γ_k, $k \in \mathbb{N}$ are such that (4.7) holds for the associated operators Π_k, we obtain

$$\begin{aligned}\|f_\infty - \mathcal{S}_L(f_\infty)\|_{L^2_\mu(\Gamma)} &= \left\| f_\infty - \left(\sum_{l=0}^{L}(f_l - f_{l-1}) - \sum_{l=0}^{L}(\mathrm{Id} - \Pi_{L-l})(f_l - f_{l-1})\right)\right\|_{L^2_\mu(\Gamma)}\\ &\leq \|f_\infty - f_L\|_{L^2_\mu(\Gamma)} + \sum_{l=0}^{L} \|\mathrm{Id} - \Pi_{L-l}\|_{F\to L^2_\mu(\Gamma)} \|f_l - f_{l-1}\|_F\\ &\lesssim \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right) + M^{-\alpha}\sum_{l=0}^{L}\exp\left(-(L-l)\frac{\alpha}{\sigma+\alpha}\right)\exp\left(-l\frac{\beta_s}{\gamma+\beta_s}\right)\\ &= \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right) + M^{-\alpha}\exp\left(-L\frac{\alpha}{\sigma+\alpha}\right)\sum_{l=0}^{L}\exp\left(l\left(\frac{\alpha}{\sigma+\alpha} - \frac{\beta_s}{\gamma+\beta_s}\right)\right),\end{aligned} \tag{4.9}$$

where we used Assumption A1. Again, we distinguish the cases (a)–(c).

(a) $\gamma/\beta_s < \sigma/\alpha$. In this case $\alpha/(\sigma+\alpha) < \beta_s/(\gamma+\beta_s)$. Thus, the sum on the right-hand side of (4.9) is uniformly bounded in L and we obtain

$$\begin{aligned}\|f_\infty - \mathcal{S}_L(f_\infty)\|_{L^2(\mu)} &\lesssim \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right) + \exp\left(-L\frac{\alpha}{\sigma+\alpha}\right)\\ &\lesssim \exp\left(-L\frac{\alpha}{\sigma+\alpha}\right),\end{aligned}$$

where we used the fact that $\beta_w \geq \beta_s$ for the last inequality.

(b) $\gamma/\beta_s = \sigma/\alpha$. In this case $\alpha/(\sigma+\alpha) = \beta_s/(\gamma+\beta_s)$. Thus, the sum on the right-hand side of (4.6) equals $L+1$ and we obtain

$$\begin{aligned}\|f_\infty - \mathcal{S}_L(f_\infty)\|_{L^2(\mu)} &\lesssim \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right) + \exp\left(-L\frac{\alpha}{\sigma+\alpha}\right)(L+1)\\ &\lesssim \exp\left(-L\frac{\alpha}{\sigma+\alpha}\right)(L+1),\end{aligned}$$

where we used the fact that $\beta_w \geq \beta_s$ for the last inequality.

(c) $\gamma/\beta_s > \sigma/\alpha$. In this case $\alpha/(\sigma+\alpha) > \beta_s/(\gamma+\beta_s)$. Thus, the sum on the right-hand side of (4.6) is a divergent geometric series and we obtain

$$\begin{aligned}\|f_\infty - \mathcal{S}_L(f_\infty)\|_{L^2(\mu)} &\lesssim \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right) + M^{-\alpha}\exp\left(-L\frac{\beta_s}{\gamma+\beta_s}\right)\\ &\lesssim \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right),\end{aligned}$$

where we used the definition of $M = \exp(L\delta)$ and δ in the case $\gamma/\beta_s > \sigma/\beta_w$ in the last inequality.

Conclusion. It remains to choose L such that the residual bound equals ϵ and insert this choice of L into the work bound. For simplicity, we assume L can be any real number. In practice, rounding up to the next largest value decreases the residual and increases the work only by a constant factor. One final time, we distinguish the cases (a)–(c).

(a) $\gamma/\beta_s < \sigma/\alpha$. Defining L as the solution of

$$\exp\left(-L\frac{\alpha}{\sigma+\alpha}\right) = \epsilon,$$

we obtain the second inequality in the following estimate:

$$\begin{aligned}\text{Work}\left(\mathcal{S}_L(f_\infty)\right) &\lesssim \exp\left(L\frac{\sigma}{\sigma+\alpha}\right)(L+1)^2\log(L+1)\\ &\lesssim \epsilon^{-\lambda}|\log\epsilon|^2\log|\log\epsilon|.\end{aligned}$$

(b) $\gamma/\beta_s = \sigma/\alpha$. Since we assumed that $\epsilon \lesssim 1$ there is a unique positive solution of

$$\exp\left(-L\frac{\alpha}{\sigma+\alpha}\right)(L+1) = \epsilon.$$

With this choice of L we obtain the second inequality in the following estimate:

$$\begin{aligned}\text{Work}\left(\mathcal{S}_L(f_\infty)\right) &\lesssim \exp\left(L\frac{\sigma}{\sigma+\alpha}\right)(L+1)^3\log(L+1)\\ &\lesssim \epsilon^{-\lambda}|\log\epsilon|^{3+\lambda}\log|\log\epsilon|.\end{aligned}$$

(c) $\gamma/\beta_s > \sigma/\alpha$. We assume $\beta_w > \beta_s$, the case $\beta_w = \beta_s$ can be treated analogously. Defining L as the solution of

$$\exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right) = \epsilon,$$

we obtain the second inequality in the following estimate:

$$\begin{aligned}\text{Work}\left(\mathcal{S}_L(f_\infty)\right) &\lesssim \exp\left(L\left(\frac{\gamma}{\gamma+\beta_s}+\sigma\delta\right)\right)(L+1)^2\log(L+1)\\ &\lesssim \epsilon^{-\lambda}|\log\epsilon|^2\log|\log\epsilon|.\end{aligned}$$

In all cases, our choice of L satisfies $L \gtrsim |\log\epsilon|$, thus $\mathbb{P}(E^c) \lesssim L^{-L} \lesssim \epsilon^{\log|\log\epsilon|}$ by (4.8). □

Remark 4.4. The proof does not exploit independence of samples across different Γ_k, $k \in \{0,\dots,L\}$, but instead relies on a simple union bound (see (4.8)). Thus, we could alternatively first create Γ_L and then define all Γ_l with $l < L$ as subsets of it.

Remark 4.5. After the functions $f_l - f_{l-1}$ have been evaluated in all $\boldsymbol{y} \in \Gamma_{L-l}$, determining the polynomial coefficients of $\Pi_{L-l}(f_l - f_{l-1})$, $l \in \{0, \dots, L\}$ with accuracy $\epsilon > 0$ requires

$$|\log \epsilon| \sum_{k=0}^{L} m_k^{2\sigma} = |\log \epsilon| \sum_{k=0}^{L} M^{2\sigma} \exp\left(2k\sigma/(\sigma+\alpha)\right) \lesssim |\log \epsilon| M^{2\sigma} \exp\left(2L\sigma/(\sigma+\alpha)\right)$$

operations. Indeed, matrix vector products with the Gramian matrices $\mathbf{G}_k$ of (2.3) require $m|\Gamma_k| = \mathcal{O}(m_k^{2\sigma})$, according to Remark 2.1. Furthermore, in the event E in which the estimate of the previous theorem holds, the condition numbers of these matrices are bounded by 3, such that suitable iterative algorithms require $\mathcal{O}(|\log \epsilon|)$ iterations to achieve accuracy $\epsilon > 0$.

Inspection of the proof of the previous theorem shows that, even if we include this cost in the work specification, the conclusion holds true with slightly different logarithmic factors and the exponent

$$\tilde{\lambda} := \begin{cases} 2\sigma/\alpha & \text{if} \quad \gamma/\beta_s \leq 2\sigma/\alpha \\ \gamma/\beta_s & \text{if} \quad \gamma/\beta_s > 2\sigma/\alpha, \end{cases}$$

instead of λ (assuming for simplicity that $\beta_s = \beta_w$), provided that we change the definition of the subsequence m_k in (4.1) to

$$m_k := \exp\left(k/(2\sigma+\alpha)\right).$$

However, in our numerical experiments, we stick to $m_k := \exp\left(k/(\sigma+\alpha)\right)$ since the cost to determine the polynomial coefficients is practically negligible.

To obtain mean square convergence, we replace the least squares approximations Π_k by the stabilized versions Π_k^c from part (iii) of Theorem 2.2, and define

$$\mathcal{S}_L^c(f_\infty) := \Pi_L^c f_0 + \sum_{l=1}^{L} \Pi_{L-l}^c (f_l - f_{l-1}). \tag{4.10}$$

Theorem 4.6 (Mean square convergence)**.** *Let $0 < \epsilon \lesssim 1$. If Assumptions A1, A2(2), and A3 hold, then we may choose $L \in \mathbb{N}$ such that*

$$\mathbb{E}\, \|f_\infty - \mathcal{S}_L^c(f_\infty)\|_{L_\mu^2(\Gamma)}^2 \leq \epsilon^2 \tag{4.11}$$

and

$$\mathrm{Work}\left(\mathcal{S}_L^c(f_\infty)\right) \lesssim \epsilon^{-\lambda} |\log \epsilon|^t \log|\log \epsilon|,$$

with λ and t as in Theorem 4.3.

Proof. The work bounds from the proof of Theorem 4.3 hold unchanged.

We next establish residual bounds for arbitrary $L \in \mathbb{N}$ as before, using the error representation

$$f_\infty - \mathcal{S}_L^c(f_\infty) = f_\infty - f_L + \sum_{l=0}^{L} (\mathrm{Id} - \Pi_{L-l}^c)(f_l - f_{l-1}).$$

The triangle inequality of the norm $(\mathbb{E}\, \|\cdot\|_{L_\mu^2(\Gamma)}^2)^{1/2}$ implies that

$$\begin{aligned} \left(\mathbb{E}\, \|f_\infty - \mathcal{S}_L^c(f_\infty)\|_{L_\mu^2(\Gamma)}^2\right)^{1/2} &\leq \left(\|f_\infty - f_L\|_{L_\mu^2(\Gamma)}^2\right)^{1/2} + \sum_{l=0}^{L} \left(\mathbb{E}\, \left\|(\mathrm{Id} - \Pi_{L-l}^c)(f_l - f_{l-1})\right\|_{L_\mu^2(\Gamma)}^2\right)^{1/2} \\ &\lesssim \|f_\infty - f_L\|_{L^2(\mu)} + \sum_{l=0}^{L} \left(e_{V_{L-l},2}^2(f_l - f_{l-1}) + \|f_l - f_{l-1}\|_{L_\mu^2(\Gamma)}^2 |\Gamma_{L-l}|^{-2\alpha/\sigma}\right)^{1/2} \\ &=: (\star) \end{aligned}$$

where we used part (iii) of Theorem 2.2 together with the fact that $L \geq 2\alpha/\sigma$ for small enough ϵ for the second inequality. We observe that

– by Assumption A1, we have

$$\|f_\infty - f_L\|_{L^2_\mu(\Gamma)} \lesssim \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right)$$

– by Assumptions A1 and A2(2), we have

$$e^2_{V_{L-l},2}(f_l - f_{l-1}) \lesssim \left(M^{-\alpha}\exp\left(-(L-l)\frac{\alpha}{\sigma+\alpha}\right)\exp\left(-l\frac{\beta_s}{\gamma+\beta_s}\right)\right)^2$$

– by (4.2)

$$\|f_l - f_{l-1}\|^2_{L^2_\mu(\Gamma)}\,|\Gamma_{L-l}|^{-2\alpha/\sigma} \lesssim \left(M^{-\alpha}\exp\left(-l\frac{\beta_w}{\gamma+\beta_s}\right)\exp\left(-(L-l)\frac{\alpha}{\sigma+\alpha}\right)\right)^2.$$

Combining these observations we arrive at

$$\begin{aligned}(\star) &\lesssim \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right) + M^{-\alpha}\sum_{l=0}^{L}\exp\left(-(L-l)\frac{\alpha}{\sigma+\alpha} - l\frac{\beta_s}{\gamma+\beta_s}\right)\\ &\lesssim \exp\left(-L\frac{\beta_w}{\gamma+\beta_s}\right) + M^{-\alpha}\exp\left(-L\frac{\alpha}{\sigma+\alpha}\right)\sum_{l=0}^{L}\exp\left(l\left(\frac{\alpha}{\sigma+\alpha} - \frac{\beta_s}{\gamma+\beta_s}\right)\right).\end{aligned}$$

From here, the proof may be concluded exactly as that of Theorem 4.3. □

5. An adaptive algorithm

We introduce in this section an adaptive algorithm for the case when an optimal sequence of polynomial subspaces, the rate of convergence $f_l \to f_\infty$, or the cost for evaluations of f_l are unknown.

To describe our algorithm, we restrict ourselves to the case when $\Gamma = [0,1]^d$, $d \in \mathbb{N}$ and when $\mu = \lambda$ is the Lebesgue measure. By the results in Section 3.2, we may then use samples and weights from the arcsine distribution instead of the optimal distributions. An alternative strategy for sampling in adaptive algorithms is presented in [1].

We start by describing the building blocks that are used by our adaptive algorithm to select polynomial approximation subspaces.

Definition 5.1 (Multivariate Legendre polynomials)**.**

(i) We denote by $(P_i)_{i\in\mathbb{N}}$ the univariate $L^2_\lambda([0,1])$-orthonormal Legendre polynomials and define their tensor products

$$P_\eta := \bigotimes_{j=1}^{d} P_{\eta_j} : [0,1]^d \to \mathbb{R},$$

$$P_\eta(\boldsymbol{y}) := \prod_{j=1}^{d} P_{\eta_j}(\boldsymbol{y}_j)$$

for $\eta \in \mathbb{N}^d$.

(ii) For each multi-index $\mathbf{k} \in \mathbb{N}^d$, we define the polynomial subspace

$$\mathcal{P}_{\mathbf{k}} := \operatorname{span}\{P_\eta : 2^{\mathbf{k}} - 1 \leq \eta < 2^{\mathbf{k}+1} - 1\} \subset L^2([0,1]^d, \lambda).$$

Remark 5.2 (Orthonormal decomposition)**.** Since polynomials are dense in $L^2_\lambda([0,1]^d)$, the subspaces $(\mathcal{P}_\mathbf{k})_{\mathbf{k}\in\mathbb{N}^d}$ form an orthonormal decomposition of $L^2_\lambda([0,1]^d)$. We use exponentially large subspaces instead of the simpler, one-dimensional subspaces $\mathcal{P}_\mathbf{k} = \mathbb{R}\cdot P_\mathbf{k}$ to avoid computational overhead resulting from slow construction of large polynomial subspaces.

We use the notation $f_{-1} := 0$ to avoid separate treatment of the term corresponding to $l = 0$ in the following. To describe a multilevel approximation, we need to construct a sequence $(V_k)_{k=0}^L$ of polynomial subspaces, such that the difference $f_l - f_{l-1}$ is projected onto V_{L-l} using weighted least squares approximation. The final approximation is then defined as

$$\sum_{l=0}^{L} \Pi_{L-l}(f_l - f_{l-1}). \tag{5.1}$$

where Π_k projects onto V_k for $0 \le k \le L$. As in Section 4, if the samples used by Π_k are distributed according to the optimal distribution of V_k, then we require that the number of samples N_k satisfy

$$\kappa \frac{N_k}{\log N_k} \ge \dim V_k \tag{5.2}$$

for some $\kappa > 0$. As an alternative, we may use samples from the arcsine distribution, which is independent of the polynomial subspaces V_k. By Section 3.2, this increases the number of required samples only by a constant factor.

To construct the sequence of polynomial subspaces in an adaptive fashion, our algorithm constructs a (finite) downward closed multi-index set $\mathcal{I} \subset \mathbb{N}^{d+1}$. Given such a set, we let

$$V_k := \bigoplus_{\mathbf{k}\in\mathbb{N}^d:(\mathbf{k},L-k)\in\mathcal{I}} \mathcal{P}_\mathbf{k} \quad 0 \le k \le L,$$

where

$$L := \max\{l \in \mathbb{N} : \exists \mathbf{k} \in \mathbb{N}^d \text{ s.t. } (\mathbf{k}, l) \in \mathcal{I}\} < \infty,$$

which means that we project the difference $f_l - f_{l-1}$ onto the subspace V_{L-l} that is determined by the slice $\mathcal{I}_l := \{\mathbf{k} \in \mathbb{N}^d : (\mathbf{k}, l) \in \mathcal{I}\}$ of the multi-index set $\mathcal{I}$. Let

$$\mathcal{A}(\mathbf{k}, l) := \{(\mathbf{k}', l') \in \mathbb{N}^{d+1} \setminus \mathcal{I} : |\mathbf{k} - \mathbf{k}'| + |l - l'| = 1 \text{ and } \mathcal{I} \cup \{\mathbf{k}, l\} \text{ is downward closed}\}$$

denote the set of admissible multi-indices that are neighbouring $(\mathbf{k}, l)$. For each multi-index $(\mathbf{k}, l) \in \mathcal{I}$, we compute the norm of the projection of $f_l - f_{l-1}$ onto $\mathcal{P}_\mathbf{k}$. This norm represents the gain that was made by adding $(\mathbf{k}, l)$ to $\mathcal{I}$. Furthermore, we estimate the work that adding this multi-index incurred. The work could be estimated directly using a timing function, or it can be based on a work model, *e.g.* the product of work per samples times number of needed samples in (5.2). With these ingredients, we can construct $\mathcal{I}$ similarly to [13, 20]. We simply start with $\mathcal{I} = \{\mathbf{0}\}$ then for every iteration of our algorithm, we find the index $(\mathbf{k}, l) \in \mathcal{I}$ which has a non-empty set of neighbouring admissible multi-indices and which maximizes the ratio between the gain and work estimates. Finally, we add those neighbours to the set $\mathcal{I}$ and repeat. Algorithm 1 gives a summary of our algorithm in pseudocode.

The adaptive algorithm can fail, for example, when there are multiple zero coefficients, which would prevent the algorithm from exploring further non-zero coefficients beyond them. We expect the algorithm to perform optimally when the coefficients decay monotonically (or are not too far from doing so) but we cannot prove this conjecture. Instead, we refer to the numerical experiments in Section 7 below, where the adaptive algorithm performs as good as the method that exploits *a priori* information.

Algorithm 1. Adaptive multilevel algorithm.

function MLA($(f_l)_{l\in\mathbb{N}}$,STEPS)
  $\mathcal{I} \leftarrow \{\mathbf{0}\}$
  $X_l \leftarrow \varnothing \ \forall l \in \mathbb{N}$
  $\Delta_l \leftarrow 0 \ \forall l \in \mathbb{N}$
  for $0 \le i <$ STEPS **do**
    $(\mathbf{k}, l) \leftarrow \arg\max_{(\mathbf{k},l)\in\mathcal{I}} \frac{\|\operatorname{Proj}_{\mathbf{k}(j)} \Delta_{l(j)}\|_{L^2_\lambda}}{\operatorname{WORK}(\mathbf{k},l)}$
    $\mathcal{L} = \{l' : (\mathbf{k}', l') \in \mathcal{A}(\mathbf{k}, l)\}$
    for $l' \in \mathcal{L}$ **do**
      $N_+ \leftarrow N(\mathcal{I}_{l'} \cup \{\mathbf{k}' : (\mathbf{k}', l') \in \mathcal{A}(\mathbf{k}, l)\}) - N(\mathcal{I}_l)$
      for $0 \le j < N_+$ **do**
        Generate $\boldsymbol{y} \sim p_d^\infty$
        $y \leftarrow (f_l - f_{l-1})(\boldsymbol{y})$
        $X_l \leftarrow X_l \cup \{(\boldsymbol{y}, y)\}$
      end for
      $\Delta_l \leftarrow \Pi_{L-l}(f_l - f_{l-1})$
    end for
    $\mathcal{I} \leftarrow \mathcal{I} \cup \mathcal{A}(\mathbf{k}, l)$
  end for
  return $\sum_{0\le l\le L} \Delta_l$
end function

6. Application to parametric PDE

We assume in this section that $u(\cdot, \boldsymbol{y})$ is the solution of some partial differential equation (PDE) with parameters $\boldsymbol{y} \in \Gamma \subset \mathbb{R}^d$ and that we are interested in the *response surface*

$$\boldsymbol{y} \mapsto f_\infty(\boldsymbol{y}) := Q(u(\cdot, \boldsymbol{y})) \in \mathbb{R},$$

where $Q(u(\cdot, \boldsymbol{y}))$ is a real-valued quantity of interest, such as a point evaluation, a spatial average, or a maximum. In most situations, we cannot evaluate $f_\infty(\boldsymbol{y})$ exactly, as this would require an analytic solution of the PDE. Instead, we have to work with discretized solutions $u_n(\cdot, \boldsymbol{y})$ for each $\boldsymbol{y}$, which yield approximate response surfaces

$$\begin{aligned} f_n \colon \Gamma &\to \mathbb{R} \\ \boldsymbol{y} &\mapsto Q(u_n(\cdot, \boldsymbol{y})). \end{aligned}$$

For example, if we employ finite element discretizations with maximal element diameter $h := n^{-1}$, then the work required for evaluations of f_n grows like $h^{-\gamma} = n^\gamma$ for some $\gamma > 0$. To apply the multilevel method of Section 4, we need to verify the remaining Assumptions A1 and A2 from there.

As a motivating example, we consider a linear elliptic second order PDE, which has been extensively studied in recent years [2, 5, 7, 19],

$$\begin{aligned} -\nabla \cdot (a(x, \boldsymbol{y}) \nabla u(x, \boldsymbol{y})) &= g(x) \quad \text{in} \quad U \subset \mathbb{R}^D \\ u(x, \boldsymbol{y}) &= 0 \quad \text{on} \quad \partial U, \end{aligned} \tag{6.1}$$

with $a : U \times \Gamma \to \mathbb{R}$ and $\Gamma := [0, 1]^d$.

Proposition 6.1. For any $n \in \mathbb{N}$, let u_n be finite element approximations of order $r \ge 1$ and maximal element diameter $h := (n + 1)^{-1}$, and let $f_n(\boldsymbol{y}) := Q(u_n(\cdot, \boldsymbol{y}))$. Assume that g and U are sufficiently smooth, that

$$\inf_{x\in U, \boldsymbol{y}\in\Gamma} a(x, \boldsymbol{y}) > 0, \tag{6.2}$$

and that Q is a continuous linear functional on $L^2(U)$.

(i) If $a \in C^r(U \times \Gamma)$ for some $r \geq 1$, then

$$\|f_\infty - f_n\|_{L^2(\Gamma)} \lesssim h^{r+1}$$

and

$$\|f_\infty - f_n\|_{C^{r-1}(\Gamma)} \lesssim h^2.$$

(ii) If for some $r, s \geq 1$ we have

$$a \in C^r(U) \otimes C^s(\Gamma) := \left\{ a \colon U \times \Gamma \to \mathbb{R} : \left\| \partial_x^{\mathbf{r}} \partial_{\boldsymbol{y}}^{\mathbf{s}} a \right\|_{C^0(U\times\Gamma)} < \infty \ \forall\, |\mathbf{r}|_1 \leq r, |\mathbf{s}|_1 \leq s \right\}, \tag{6.3}$$

then

$$\|f_\infty - f_n\|_{C^s(\Gamma)} \lesssim h^{r+1}.$$

Proof. In both cases, the standard theory of second order elliptic differential equations shows that $\boldsymbol{y} \mapsto u(\cdot, \boldsymbol{y})$ is well defined as a map from Γ into $H^{r+1}(U)$, with

$$\|u\|_{L^\infty(\Gamma; H^{r+1}(U))} < \infty.$$

Next, we observe that the derivatives $\partial_{\boldsymbol{y}_j} u(\cdot, \boldsymbol{y})$, $j \in \{1, \ldots, d\}$ satisfy PDEs with the same operator as in (6.1) but with new right-hand sides

$$\tilde{g}(x) := \nabla \cdot (\partial_{\boldsymbol{y}_j} a(x, \boldsymbol{y}) \nabla u(x, \boldsymbol{y})).$$

The regularity of this right-hand side now depends on the assumptions on the coefficient a. In case (i) we have $\partial_{\boldsymbol{y}_j} a(\cdot, \boldsymbol{y}) \in C^{r-1}(U)$ and thus $\tilde{g} \in H^{r-2}(U)$. Therefore, $\partial_{\boldsymbol{y}_j} u(\cdot, \boldsymbol{y}) \in H^r(U)$ for each $\boldsymbol{y} \in \Gamma$ and, moreover, we have the uniform estimate

$$\left\| \partial_{\boldsymbol{y}_j} u \right\|_{L^\infty(\Gamma; H^r(U))} < \infty.$$

In case (ii) we have $\partial_{\boldsymbol{y}_j} a(\cdot, \boldsymbol{y}) \in C^r(U)$ and thus $\tilde{g} \in H^{r-1}(U)$. Therefore, $\partial_{\boldsymbol{y}_j} u(\cdot, \boldsymbol{y}) \in H^{r+1}(U)$ for each $\boldsymbol{y} \in \Gamma$ and, moreover, we have the uniform estimate

$$\left\| \partial_{\boldsymbol{y}_j} u \right\|_{L^\infty(\Gamma; H^{r+1}(U))} < \infty.$$

Repeatedly applying these arguments yields

$$\|u\|_{C^{r-1}(\Gamma; H^2(U))} < \infty,$$

and

$$\|u\|_{C^s(\Gamma; H^{r+1}(U))} < \infty,$$

in cases (i) and (ii), respectively. We may now conclude by using standard finite-element theory. In case (i), we have

$$\begin{aligned} \|f_\infty - f_n\|_{L^2(\Gamma)} &\leq \|Q\| \, \|u - u_n\|_{L^2(\Gamma; L^2(U))} \\ &\lesssim h^{r+1} \, \|u\|_{L^2(\Gamma; H^{r+1}(U))} \end{aligned}$$

and

$$\begin{aligned} \|f_\infty - f_n\|_{C^{r-1}(\Gamma)} &\lesssim \|u - u_n\|_{C^{r-1}(\Gamma; L^2(U))} \\ &\lesssim h^2 \, \|u\|_{C^{r-1}(\Gamma; H^2(U))}, \end{aligned}$$

whereas in case (ii), we have

$$\begin{aligned} \|f_\infty - f_n\|_{C^s(\Gamma)} &\lesssim \|u - u_n\|_{C^s(\Gamma; L^2(U))} \\ &\lesssim h^{r+1} \, \|u\|_{C^s(\Gamma; H^{r+1}(U))}, \end{aligned}$$

□

Remark 6.2. In case (i) of the previous proposition, differentiating with respect to $\boldsymbol{y}$ reduces the number of available derivatives in x, which are required for convergence of the finite element method. Thus, the convergence in $L^2(\Gamma)$ is faster than that in $C^{r-1}(\Gamma)$. Case (ii), on the other hand, describes the so-called *mixed smoothness* of the coefficient in x and $\boldsymbol{y}$, meaning that differentiating in $\boldsymbol{y}$ does not affect the differentiability with respect to x.

If the coefficients depend analytically on $\boldsymbol{y}$, then the same holds for f_∞, which can be exploited to obtain algebraic polynomial approximability rates of f_∞ even in the case of infinite-dimensional parameters [5,16], as shown below.

Proposition 6.3. Let $\Gamma := [-1,1]^\infty$. Assume that Q is a linear and continuous functional on $L^2(U)$, that $0 < \inf_{x,\boldsymbol{y}} a(x,\boldsymbol{y}) \leq \sup_{x,\boldsymbol{y}} a(x,\boldsymbol{y}) < \infty$, and that

$$a(x,\boldsymbol{y}) = \bar{a}(x) + \sum_{j=0}^{\infty} y_j \psi_j(x),$$

$$a(x,\boldsymbol{y}) = \bar{a}(x) + \left(\sum_{j=0}^{\infty} y_j \psi_j(x)\right)^2,$$

or

$$a(x,\boldsymbol{y}) = \exp\left(\sum_{j=0}^{\infty} y_j \psi_j(x)\right).$$

If there exists $r_{\max} > 1$ such that

$$\|\psi_j\|_{C^r(U)} \lesssim (j+1)^{-(r_{\max}+1-r)} \qquad \forall j \in \mathbb{N},\ 0 \leq r < r_{\max},$$

then, for any $r \in \mathbb{N}$ with $1 \leq r < r_{\max}$, finite element approximations with maximal element diameter $h := (n+1)^{-1}$ achieve

$$\|f_\infty - f_n\|_{L^\infty(\Gamma)} \leq C h^{r+1}$$

with a constant C independent of n. Furthermore, for any such r, there is a sequence $(V_m)_{m\in\mathbb{N}}$ of downward closed polynomial spaces with $\dim V_m = m$ such that finite element approximations with order r and maximal diameter $h := (n+1)^{-1}$ achieve

$$e_{V_m,1,\infty}(f_\infty - f_n) \leq C(m+1)^{-\alpha} h^{r+1} \qquad \forall\, 0 < \alpha < r_{\max} - r$$

with a constant C independent of n and m.

Proof. It was shown in Theorem 4.1 and Section 5 of [5] that for each $0 \leq r < r_{\max}$ there exists a set $\Gamma_r \subset \mathbb{C}^\infty$, $\Gamma \subset \Gamma_r$ such that $\|a\|_{L^\infty(\Gamma_r;C^r(U))} < \infty$ and such that $\boldsymbol{y} \mapsto u(\cdot,\boldsymbol{y})$ may be extended to a complex differentiable map from Γ_r into $H^{1+r}(U)$ with

$$\|u\|_{L^\infty(\Gamma_r;H^{1+r}(U))} < \infty. \tag{6.4}$$

For a detailed description of the sets Γ_r we refer to [5]. For our purposes it suffices to know that the better the summability of $(\|\psi_j\|_{C^r(U)})_{j\in\mathbb{N}}$, the larger Γ_r can be chosen; and the larger Γ_r the better the polynomial approximability properties of complex differentiable maps defined on Γ_r. In particular, the results of Section 2 of [5], show that when restricted to the smaller set Γ such maps may be approximated at algebraic convergence rates within downward closed polynomial subspaces. More specifically, equation (2.27) of [5] shows that if a function

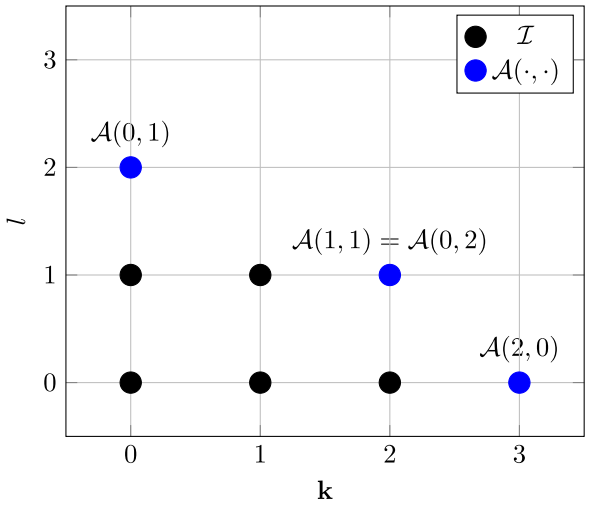


FIGURE 1. Example with $d = 1$ of a multi-index set $\mathcal{I}$ and the associated non-empty sets of neighbouring admissible multi-indices $\mathcal{A}(\cdot,\cdot)$. In this example $L = 1$, $V_1 = \operatorname{span}\{1, \boldsymbol{y}, \dots, \boldsymbol{y}^6\} = \operatorname{span}\{P_0(\boldsymbol{y}), \dots, P_6(\boldsymbol{y})\}$, and $V_0 = \operatorname{span}\{1, \boldsymbol{y}, \boldsymbol{y}^2\} = \operatorname{span}\{P_0(\boldsymbol{y}), P_1(\boldsymbol{y}), P_2(\boldsymbol{y})\}$.

e is complex differentiable on Γ_r, then for any $m \in \mathbb{N}$ there exists a downward closed polynomial subspace V_m such that

$$\inf_{\tilde{v} \in V_m \otimes L^2(U)} \|e - \tilde{v}\|_{L^\infty(\Gamma; L^2(U))} \lesssim (m+1)^{-\alpha} \|e\|_{L^\infty(\Gamma_r; L^2(U))}$$

for all $\alpha < r_{\max} - r$. Applying this estimate with $e := u - u_n$ shows

$$\begin{aligned} \inf_{v \in V_m} \|(f_\infty - f_n) - v\|_{L^\infty(\Gamma)} &\le \|Q\| \inf_{\tilde{v} \in V_m \otimes L^2(U)} \|(u - u_n) - \tilde{v}\|_{L^\infty(\Gamma; L^2(U))} \\ &\lesssim (m+1)^{-\alpha} \|u - u_n\|_{L^\infty(\Gamma_r; L^2(U))} . \end{aligned}$$

By standard finite element analysis we finally obtain

$$\|u - u_n\|_{L^\infty(\Gamma_r; L^2(U))} \le C h^{r+1} \|u\|_{L^\infty(\Gamma_r; H^{r+1}(U))} .$$

with $C = C\big(\|a\|_{L^\infty(\Gamma_r; C^r(U))}\big) < \infty$. Combining the previous two estimates with (6.4) concludes the proof. □

Remark 6.4. Similar results can also be shown for PDEs of parabolic type and for some nonlinear PDEs [5].

7. NUMERICAL EXPERIMENTS

To support our theoretical analysis, we performed numerical experiments on linear elliptic parametric PDEs of the form

$$\begin{aligned} -\nabla \cdot (a(x, \boldsymbol{y}) \nabla u(x, \boldsymbol{y})) &= 1 \quad \text{in} \quad U := [-1, 1]^D \\ u(x, \boldsymbol{y}) &= 0 \quad \text{on} \quad \partial U, \end{aligned} \tag{7.1}$$

as in Section 6.

We let

$$a(x, \boldsymbol{y}) = 1 + \|x\|_2^r + \|\boldsymbol{y}\|_2^s, \quad \boldsymbol{y} \in \Gamma := [-1, 1]^d$$

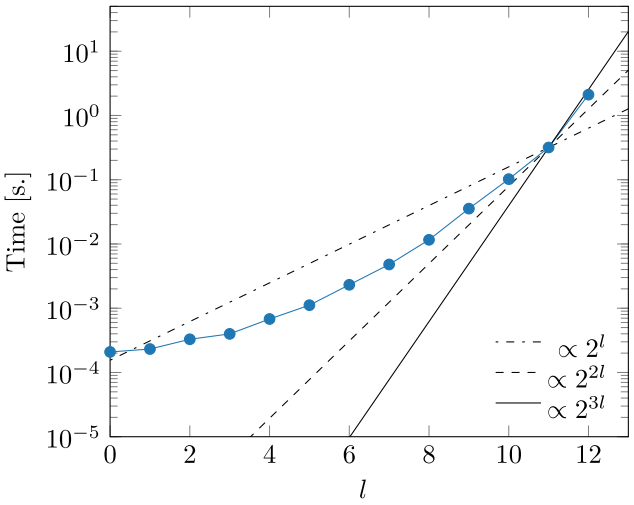


FIGURE 2. The running time of computing a single sample of (7.2) with $d = 6$ using a discretisation with a constant mesh size $h_l = 2^{-l}/8$. In theory, the work should grow like $\mathcal{O}(h_l^{-3})$, *i.e*, $\mathcal{O}(2^{3l})$. However, for most levels, the work grows on average like $\mathcal{O}(2^{2l})$, hence we choose $\gamma = 2$ in our numerical tests and discussion. Note that level 12 does not follow the model $\mathcal{O}(2^{2l})$.

for $r := 1$, $s := 3$, $D := 2$ and $d \in \{2, 3, 4, 6\}$. Our goal was to approximate the response surface

$$\boldsymbol{y} \mapsto f(\boldsymbol{y}) := Q(u(\cdot, \boldsymbol{y})) := \int_U u(\cdot, \boldsymbol{y}) \, \mathrm{d}x \tag{7.2}$$

in $L^2(\Gamma)$.

The numerical scheme we used to solve (7.1) employs centered finite difference approximations of the derivatives with a constant mesh size, h_l, for the discretization level, l, and a GMRES solver. Such a numerical scheme converges asymptotically at a rate of $\mathcal{O}(h_l^2)$ in the L^2 norm and requires a computational work of $\mathcal{O}(h_l^{-3})$, since the PDE is two-dimensional and we are using GMRES. This corresponds to the values $\beta_s = \beta_w = 2$ and $\gamma = 3$ for the parameters in Assumptions A2 and A3. However, we noticed that $\gamma = 2$ is a better fit for most discretization levels that we use, for $h_l = 2^{-l}/8$; see Figure 2. Hence, we fix $\gamma = 2$ in our tests and the discussion below.

To estimate the projection error of our estimate we evaluate the L^2 error norm using Monte Carlo,

$$\|f - S_L(f)\|^2_{L^2(\Gamma)} \approx \frac{1}{M} \sum_{j=1}^{M} (f_{L+1}(\boldsymbol{y}_j) - S_L(f)(\boldsymbol{y}_j))^2. \tag{7.3}$$

The number of samples, M, is chosen such that the estimated error of the Monte Carlo approximation is less than 10% of the norm that is approximated. In our tests we employ both the non-adaptive and the adaptive algorithms from Sections 4 and 5 with the random points being sampled from both the arcsine and the optimal distribution (using Acceptance/Rejection sampling for the latter). We also consider using the arcsine distributions with the non-adaptive algorithm. As a basis for the non-adaptive algorithm, we use total degree polynomial spaces $V_m := \operatorname{span}\{P_\eta : |\eta|_1 \leq m\}$, where P_η is a tensor product of Legendre polynomials as in Section 5. We also compare the multilevel algorithm to the straightforward, single-level approach, which for a given polynomial approximation space V_m uses samples from a fixed PDE discretization level that matches the accuracy of the polynomial best approximation in V_m. To find these matching PDE discretization levels, we consider the

complexity curve of the single-level method as the lower envelope of complexity curves with different PDE discretization levels. Even though such a method is not practical, the choice of discretization level for a given tolerance is always optimal.

Before presenting the numerical results, let us derive some *a priori* estimates of the complexity of the single-level and multilevel projection methods. From Proposition 6.1, if $a \in C^r(U) \otimes C^s(\Gamma)$, then using finite elements of order r and mesh size h would yield convergence in the space $F := C^s(\Gamma)$ with the values $\beta_s = \beta_w = r+1$ of the parameters in Section 4, and optimal solvers would require the work $\mathcal{O}(h^{-\gamma})$, $\gamma := D$. Furthermore, since functions in $C^s(\Gamma)$ are approximable by polynomials of total degree less than or equal to k at the rate $\mathcal{O}(k^{-s})$ in the supremum norm [3], we expect at least $\alpha = s$. Even though our choice $a(x, \boldsymbol{y}) = 1 + \|x\|_2^r + \|\boldsymbol{y}\|_2^s$ satisfies only $a \in C^{r-1,1}(U) \otimes C^{s-1,1}(\Gamma)$, we do not expect different rates than those derived above for $a \in C^r(U) \otimes C^s(\Gamma)$. Finally, the dimension of total degree polynomial spaces V_m equals $\binom{m+d}{d}$ and asymptotically we have $\binom{m+d}{d} \lesssim m^d$, *i.e.* $\sigma = d$.

Thus, we expect the complexity of the single-level method to be $\mathcal{O}\left(\epsilon^{-\frac{D}{r+1}-\frac{d}{s}} \log(\epsilon^{-1})\right)$, while the complexity of the multilevel method is of $\mathcal{O}\left(\epsilon^{-\max\left(\frac{D}{r+1},\frac{d}{s}\right)} \log(\epsilon^{-1})^t\right)$, where

$$t = \begin{cases} 1 & \frac{D}{r+1} > \frac{d}{s}, \\ 3 + \frac{D}{r+1} & \frac{D}{r+1} = \frac{d}{s}, \\ 2 & \frac{D}{r+1} < \frac{d}{s}. \end{cases}$$

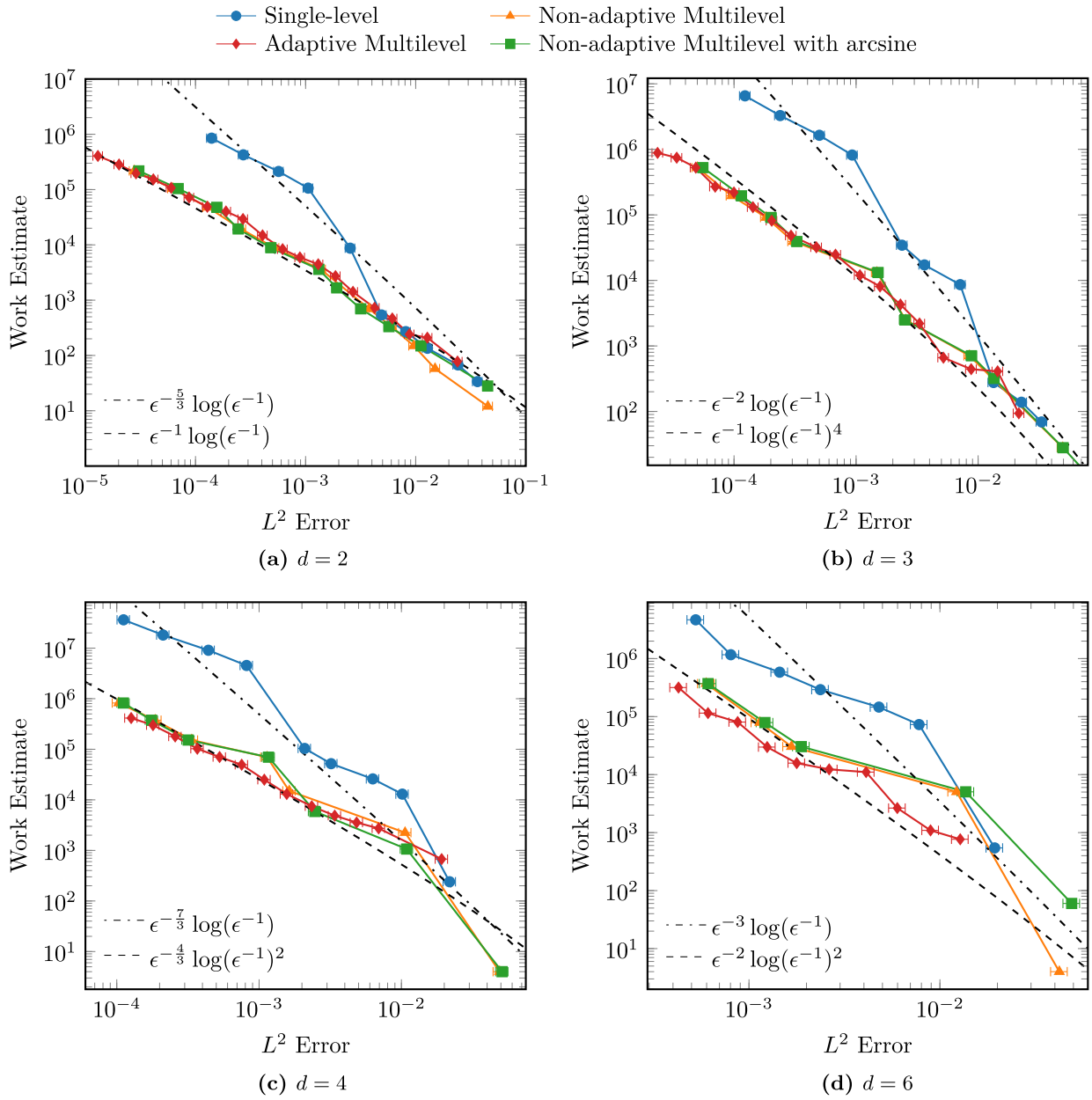
Hence, for $r = 1$ and $s = 3$, the complexity of the single-level method is $\mathcal{O}\left(\epsilon^{-1-\frac{d}{3}} \log(\epsilon^{-1})\right)$ and the complexity of the multilevel method is $\mathcal{O}\left(\epsilon^{-\max(1,\frac{d}{3})} \log(\epsilon^{-1})^t\right)$ where

$$t = \begin{cases} 1, & d < 3, \\ 4, & d = 3, \\ 2, & d > 3. \end{cases}$$

Figure 3 shows the work estimate as defined in (4.4) *vs.* the L^2 error approximation in (7.3). The theoretical rates satisfactorily match the obtained numerical rates, which show an improvement of the multilevel methods over the single-level method. They also show that sampling the random points from the arcsine distribution does not have a significant overhead compared to sampling these points from the optimal distribution. Note that the work estimate does *not* include the cost of sampling random points, the cost of assembling the projection matrix and computing the projection nor does it include the cost of finding the set $\mathcal{I}$ for the adaptive algorithm. On the other hand, Figure 4 shows the total running time in seconds of the four different methods, *including* the cost of generating points, the cost of assembling the projection matrix and computing the projection but not including the cost of finding the optimal set $\mathcal{I}$ for the adaptive algorithm. While Figure 4 still show the same complexity rates as Figure 3 for all the methods, there is a small discrepancy with the theory for $d = 6$ and very small tolerances. This is due to the fact that for small tolerances, the discretization level $l = 12$, whose work does not adequately follow the work model with $\gamma = 2$, is employed; see Figure 2.

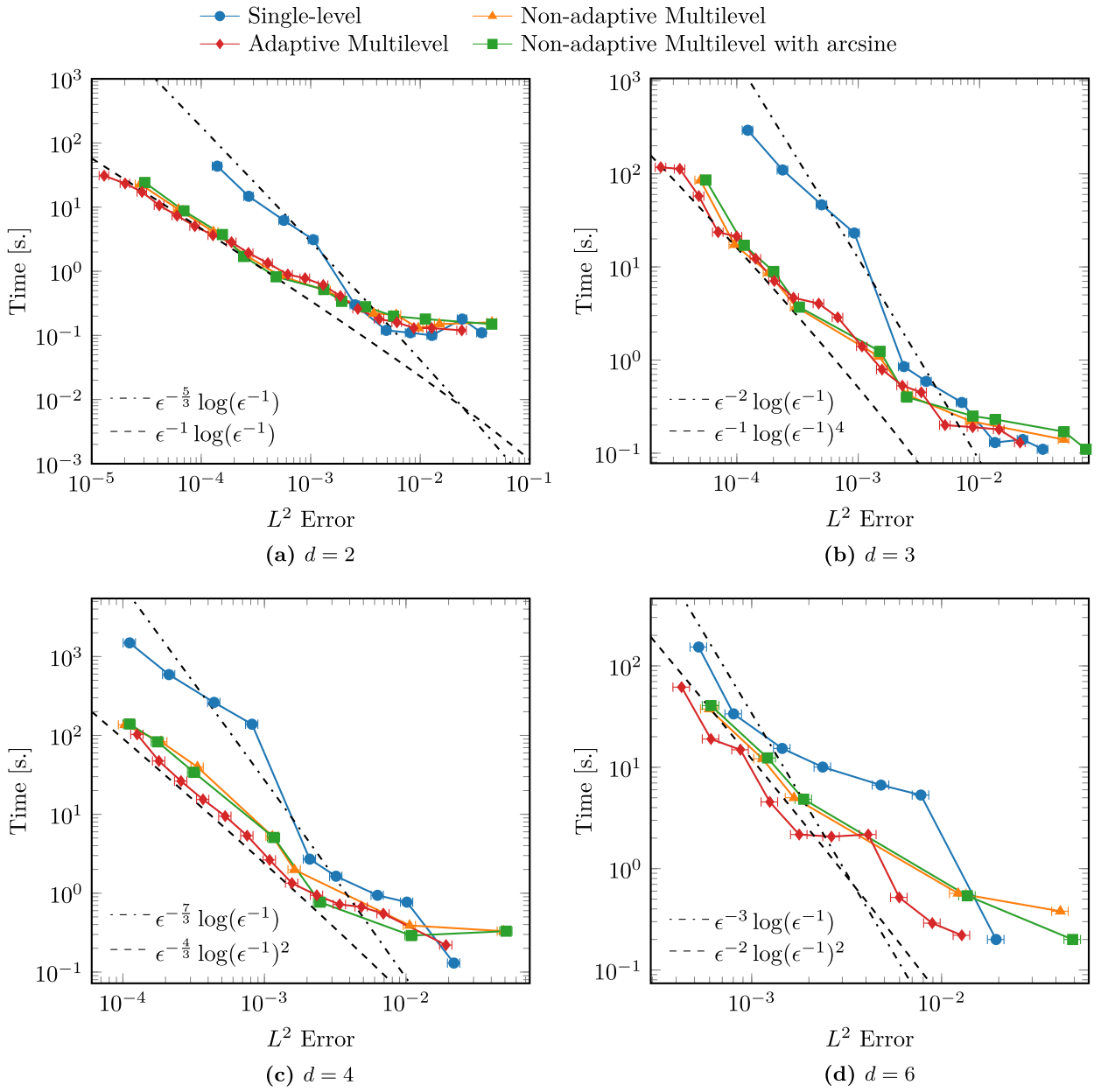
8. Conclusion

We have presented a novel multilevel projection method for the approximation of response surfaces using multivariate polynomials and random samples with different accuracies. For this purpose, we have discussed and analyzed various sampling methods for the underlying single-level approximation method. We have then presented theoretical and numerical results on our multilevel projection method for problems in which samples can be obtained at different accuracies. The numerical results show good agreement with the computational gains predicted by our theory. Future work will address the application to problems in uncertainty quantification with infinite-dimensional parameter domains and multi- or infinite-dimensional quantities of interest.



(a) $d = 2$ **(b)** $d = 3$ **(c)** $d = 4$ **(d)** $d = 6$

FIGURE 3. $L^2([-1,1]^d)$-error, approximated using (7.3) *vs.* work estimate (4.4) of single-level, non-adaptive multilevel and adaptive multilevel methods for a linear elliptic PDE with non-smooth parameter dependence. We also show the work estimate for a non-adaptive multilevel method with random points sampled from the arcsine distribution. This figure shows the agreement of the numerical results with the theoretical rates. It also shows that using the arcsine distribution does not have a significant overhead compared to using the optimal distribution for the random points.



(a) $d = 2$ **(b)** $d = 3$ **(c)** $d = 4$ **(d)** $d = 6$

FIGURE 4. Similar to Figure 3, but showing the total running time of the methods instead of their work estimate. The discrepancy with the theory for $d = 6$ and small tolerances is due to the non-asymptotic behaviour of the work-per-sample, as seen in Figure 2. Note that for small tolerances, level 12, whose work does not follow the work model with $\gamma = 2$, is employed.

Acknowledgements. F. Nobile received support from the Center for ADvanced MOdeling Science (CADMOS). R. Tempone and S. Wolfers are members of the KAUST SRI Center for Uncertainty Quantification in Computational Science and Engineering. R. Tempone received support from the KAUST CRG3 Award Ref:2281, the KAUST CRG4 Award Ref:2584, and the Alexander von Humboldt foundation. We thank an anonymous referee for their help in improving Proposition 3.3.

References

[1] B. Arras, M. Bachmayr and A. Cohen, Sequential sampling for optimal weighted least squares approximations in hierarchical spaces. Preprint arXiv:1805.10801 (2018).
[2] I.M. Babuška, R. Tempone and G.E. Zouraris, Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.* **42** (2004) 800–825.
[3] T. Bagby, L. Bos and N. Levenberg, Multivariate simultaneous approximation. *Constr. Approx.* **18** (2002) 569.
[4] A. Chkifa, A. Cohen, G. Migliorati, F. Nobile and R. Tempone, Discrete least squares polynomial approximation with random evaluations – application to parametric and stochastic elliptic PDEs. *ESAIM: M2AN* **49** (2015) 815–837.
[5] A. Chkifa, A. Cohen and C. Schwab, Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *J. Math. Pures Appl.* **103** (2015) 400–428.
[6] A. Cohen and G. Migliorati, Optimal weighted least-squares methods. Preprint arXiv:1608.00512 (2016).
[7] A. Cohen, R. Devore and C. Schwab, Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs. *Anal. App.* **9** (2011) 11–47.
[8] A. Cohen, M.A. Davenport and D. Leviatan, On the stability and accuracy of least squares approximations. *Found. Comput. Math.* **13** (2013) 819–834.
[9] M.K. Deb, I.M. Babuška and J. Tinsley Oden, Solution of stochastic partial differential equations using Galerkin finite element techniques. *Comput. Methods Appl. Mech. Eng.* **190** (2001) 6359–6372.
[10] R.A. DeVore, Nonlinear approximation. *Acta Numer.* **7** (1998) 51–150.
[11] D. Dũng, V.N. Temlyakov and T. Ullrich, Hyperbolic cross approximation. Preprint arXiv:1601.03978 (2016).
[12] J.E. Gentle, Random number generation and Monte Carlo methods, 2nd edition. In: Statistics and Computing. Springer, New York (2003).
[13] T. Gerstner and M. Griebel, Dimension–adaptive tensor–product quadrature. *Computing* **71** (2003) 65–87.
[14] M.B. Giles, Multilevel monte carlo path simulation. *Oper. Res.* **56** (2008) 607–617.
[15] M. Griebel and C. Rieger, Reproducing kernel Hilbert spaces for parametric partial differential equations. *SIAM/ASA J. Uncertainty Quant.* **5** (2017) 111–137.
[16] A.-L. Haji-Ali, F. Nobile, L. Tamellini and R. Tempone, Multi-index stochastic collocation convergence rates for random PDEs with parametric regularity. *Found. Comput. Math.* **16** (2016) 1555–1605.
[17] A.-L. Haji-Ali, F. Nobile, L. Tamellini and R. Tempone, Multi-index stochastic collocation for random PDEs. *Comput. Methods Appl. Mech. Eng.* **306** (2016) 95–122.
[18] J. Hampton and A. Doostan, Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression. *Comput. Methods Appl. Mech. Eng.* **290** (2015) 73–97.
[19] H. Harbrecht, M. Peters and M. Siebenmorgen, Multilevel accelerated quadrature for PDEs with log-normally distributed diffusion coefficient. *SIAM/ASA J. Uncertainty Quant.* **4** (2016) 520–551.
[20] M. Hegland, Adaptive sparse grids. *ANZIAM J.* **44** (2003) 335–353.
[21] S. Heinrich, Multilevel Monte Carlo methods. In: International Conference on Large-Scale Scientific Computing. Springer (2001) 58–67.
[22] F. Kuo, R. Scheichl, C. Schwab, I. Sloan and E. Ullmann, Multilevel quasi-Monte Carlo methods for lognormal diffusion problems. *Math. Comput.* **86** (2017) 2827–2860.
[23] O. Le Maître and O. Knio, Spectral Methods for Uncertainty Quantification. Springer (2010).
[24] E. Levin and D.S. Lubinsky, Christoffel functions, orthogonal polynomials, and Nevai's conjecture for Freud weights. *Constr. Approx.* **8** (1992) 463–535.
[25] J.S. Liu, Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Stat. Comput.* **6** (1996) 113–119.
[26] J.S. Liu, Monte Carlo Strategies in Scientific Computing. Springer Science & Business Media (2008).
[27] G. Mastroianni and V. Totik, Weighted polynomial inequalities with doubling and a_∞ weights. *Constr. Approx.* **16** (2000) 37–71.
[28] G. Migliorati, F. Nobile and R. Tempone, Convergence estimates in probability and in expectation for discrete least squares with noisy evaluations at random points. *J. Multivariate Anal.* **142** (2015) 167–182.
[29] A. Narayan, J. Jakeman and T. Zhou, A Christoffel function weighted least squares algorithm for collocation approximations. *Math. Comput.* **86** (2017) 1913–1947.
[30] P. Nevai, T. Erdélyi and A.P. Magnus, Generalized Jacobi weights, Christoffel functions, and Jacobi polynomials. *SIAM J. Math. Anal.* **25** (1994) 602–614.
[31] F. Nobile, R. Tempone and S. Wolfers, Sparse approximation of multilinear problems with applications to kernel-based methods in UQ. *Numer. Math.* **139** (2018) 247–280.

[32] A. Quarteroni, Some results of Bernstein and Jackson type for polynomial approximation in L^p-spaces. *Jpn J. Appl. Math.* **1** (1984) 173–181.

[33] G. Szegő, Orthogonal polynomials, 4th edition. In: Vol. XXIII of *American Mathematical Society, Colloquium Publications*. American Mathematical Society, Providence, RI (1975).

[34] J.A. Tropp, User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** (2012) 389–434.